Gaussian-Spherical Restricted Boltzmann Machines

Aurélien Decelle[1], **Cyril Furtlehner**[2]

[1] LRI-Université Paris-Sud

[2] TAU team, Inria Saclay

Abstract. We consider a special type of Restricted Boltzmann machine (RBM), namely a Gaussian-spherical RBM where the visible units have Gaussian priors while the vector of hidden variables is constrained to stay on an $\mathbb{L}_2$ sphere. The spherical constraint having the advantage to admit exact asymptotic treatments, various scaling regimes are explicitly identified based solely on the spectral properties of the coupling matrix (also called weight matrix of the RBM). Incidentally these happen to be formally related to similar scaling behaviours obtained in a different context dealing with spatial condensation of zero range processes. More specifically, when the spectrum of the coupling matrix is doubly degenerated an exact treatment can be proposed to deal with finite size effects. Interestingly the known parallel between the ferromagnetic transition of the spherical model and the Bose-Einstein condensation can be made explicit in that case. More importantly this gives us the ability to extract all needed response functions with arbitrary precision for the training algorithm of the RBM. This allows us then to numerically integrate the dynamics of the spectrum of the weight matrix during learning in a precise way. This dynamics reveals in particular a sequential emergence of modes from the Marchenko-Pastur bulk of singular vectors of the coupling matrix.

1 Introduction

In the last decade, the field of machine learning became the center of attention of both the public domain and of the scientific research. With the development of deep neural networks taking advantage of the GPU technology, the performance on classification tasks started to outperform human level at image recognition, and more recently, generative model such as generative adversarial network [1] (GAN) have been able to generate images that cannot be distinguish from a true one [2]. Despite recent significant advances [3, 4] the theoretical understanding of deep learning lag behind these progresses, in various respects like for instance on the interplay between adequate network architecture and complexity of the data.

Statistical physics has been helpful in the past to clarify the learning process on idealized inference problems. In the 80', before the A.I. winter, many works on neural networks were proposing some elements of understanding in terms of the theoretical phase diagram of some models. For instance, a retrieval phase for the Hopfield model was determined along with the number of patterns that can be retrieven in that case [5, 6, 7, 8]. Another example deals

with the perceptron where again, the capacity for storing synthetic dataset can be computed [9, 10, 11]. The storage of information in layered neural networks was also analyzed in [12] with mean-field techniques. These approaches could then be adapted in many different contexts, such as community detection [13], compressed sensing [14] or traffic inference [15] to mention only a few of them. Typically in this kind of approach, the formalism of statistical physics relates the behaviour of the model to its position on a phase diagram in the large N limit, mean-field equations being used to characterize the free energy landscape and to sample efficiently the system.

In this work a somewhat similar path is followed to study generative models, by focusing on a "tractable" version of the restricted Boltzmann machine (RBM). While RBM is considered to be a basic tool of machine learning, introduced more than 30 years ago [16], it is still attracting a lot of interest, both from the machine learning and the statistical physics communities. First, it is a model that can be handled without the need of GPU and can be run on a ordinary computer in reasonable time while solving non-trivial tasks. Second, it has only one hidden layer in its classical formulation which allows the possibility to get some understanding of the learned hidden features since they are directly linked to the visible variables. Finally, it can be expressed as an Ising model and therefore, many standard tools developed by the statistical physics community can be used to determine its properties.

Originally, the RBM played an important role in deep learning as a way to pre-train deep auto-encoders layerwise [17]. It is also in principle possible to stack many RBM to form a multi-layer generative model known as a Deep Boltzmann Machine (DBM) [18]. Within the recent years, RBM has continuously attracted the interest of the research community, firstly because it can be easily used for both continuous and discrete variables [19, 20, 21, 22] and the activation can be tuned to be either binary of relu [23]; secondly because for datasets of modest size it is able to deliver good results [24, 25] comparable to the ones obtain from more elaborated network such as GAN (see for instance [26]). However, even for such a simple model, the learning procedure (what is learned, and how it is learned) is still very difficult to analyze with non-linear activation functions, in order to identify the key features and mechanisms allowing it to work properly. Even for practical purpose, it is intrinsically difficult to efficiently estimate numerically the gradient w.r.t. the parameters of the model, as soon as the network has learned non trivial modes. Empirical procedures have been proposed, first the contrastive divergence [27] (CD) which has properties have been analyzed in [28], or the refined Persistence CD [29] (PCD) and later on a mean-field estimate [30]. None of these being fully satisfactory (see e.g. [31] for a more detailed discussion), especially if one is willing to learn an empirical distribution with good accuracy. For that purpose, recent works [32, 33] using the analogy between the RBM and the Hopfield model characterize the retrieval capacity of RBMs. RBM with sparse weight matrix have been considered in [34] to analyze compositional mechanisms of features to create complex patterns. Other works have focused on a mean-field theory for the RBMs, first to approximate the gradient and second to probe the mean-field landscape in the general case [30, 35] or in the spherical

case [36] or even to compute the entropy in a very simple case [37]. In [28] was shown that large principal components of the data were extracted in the Gaussian RBM, while the Gaussian-Bernoulli case could reveal independent components of the data. Recently we also used a mean-field approach to understand the phase diagram of the binary-binary RBM, as a function of the spectrum of the weight matrix [38, 39]. We characterized in some way how the singular modes of the weight matrix evolve and interact during learning, bringing forward a clustering model interpretation of the RBM in terms of mean-field fixed points.

In this paper we study an RBM with continuous symmetry, consisting of one layer of Gaussian variables (the visible one) and one layer of real variables with a spherical constraint. In the spirit of the original spherical Ising model introduced by Berlin and Kac [40], this offers the possibility to say something relevant to the original model, by solving a simpler one. It turns out that in a special setting the thermodynamical properties of the Gaussian-spherical RBM can be obtained exactly. This allows one to devise an exact gradient ascent of the likelihood to learn the model, despite the fact that this model as we shall see is able to encode only rather specific data. The observation made previously on the spectral dynamics of the learning procedure [38, 39], in particular that the modes of the weight matrix are learned by order of importance, will be illustrated by an exact integration of the dynamical equations introduced in these works. In addition this solution could constitute a possibility to assess approximate mean-field methods and empirical learning strategies.

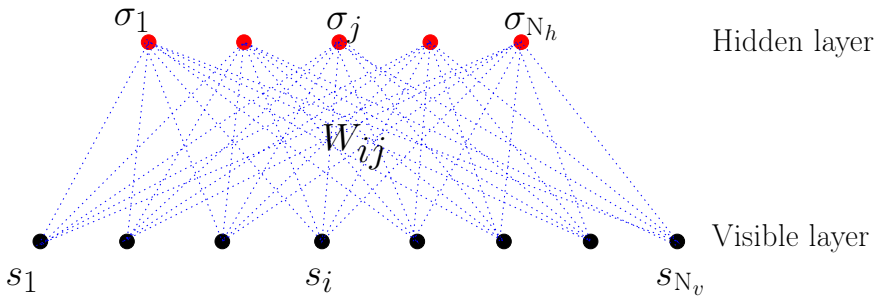

Figure 1. bipartite structure of the RBM.

The paper is organized as follows. In Section 2 we define the RBM with a spherical hidden layer and derive the likelihood and the response functions exactly in the case of Gaussian visible units, together with the dominant behavior in the thermodynamic limit. In Section 3 we specify some properties specific to the spherical constraint, as the way is occuring the onset of ferromagnetic order, the critical behavior of the magnetization associated to mode condensation, remarking and exploiting in passing some connection with spatial condensation in particle processes explored in [41]. Next, the Section 4 focuses on a particular case where the spectrum of the weight matrix is doubly-degenerated and allows one to compute exactly for finite size systems the partition function of the system. Finally, in the section 5 we exploit these results to numerically integrate the spectral dynamics of the weight matrix during learning.

2 Model definition

2.1 Boltzmann measure and associated likelihood

The basic structure of the RBM is shown on Figure 1. It is a bipartite model connecting one layer of visible variables to one layer of hidden variables, these ones acting as a field to generate interactions among visible variables. We define the visible variables $\{s_i\}_{i=1,\dots,N_v}$ and the hidden variables $\{\sigma_i\}_{i=1,\dots,N_h}$ both real valued, where N_v and N_h denotes the number of visible (resp. hidden) variables. $L = \sqrt{N_v N_h}$ will represent the size of the system and $\kappa = \sqrt{\frac{N_h}{N_v}}$ its shape. We define the energy function by

$$E(\boldsymbol{s}, \boldsymbol{\sigma}) = -\sum_{i,j} w_{ij} s_i \sigma_j + \sum_i \frac{s_i^2}{2} - \sum_i \eta_i s_i - \sum_j \theta_j \sigma_j, \tag{1}$$

W is the weight matrix between the visible and hidden variables, $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are local fields exerted on variables. In this form the visible variables have a Gaussian prior $\mathcal{N}(0,1)$ in absence of hidden variables. The spherical constraint imposes an additional prior distribution on the hidden variables. Overall the distribution over $\boldsymbol{s}$ and $\boldsymbol{\sigma}$ is defined as

$$p(\boldsymbol{s}, \boldsymbol{\sigma}) = \frac{1}{Z} e^{-E(\boldsymbol{s},\boldsymbol{\sigma})}\ \delta\big(\sum_j \sigma_j^2 - \bar{\sigma}^2 L\big) \tag{2}$$

where Z is the normalization factor and $\bar{\sigma}$ a parameter of the model. In this setting, it is possible to diagonalize the distribution by using the singular value decomposition (SVD) of the matrix W:

$$w_{ij} = \sum_\alpha w_\alpha u_i^\alpha v_j^\alpha$$

where u^α are left singular vectors, attached to the visible space, while v^α are right singular vectors attached to the hidden space and w_α are the singular values. Depending on whether $N_v > N_h$ or $N_v < N_h$ the set u^α or the set v^α is not a complete orthonormal set of respectively the visible or the hidden space. If we assume for instance that $N_h < N_v$ the matrix corresponding to the left singular vectors has to be complemented by $N_v - N_h$ arbitrary orthonormal vectors to form a complete basis of the visible space. For the moment we don't need to specify whether N_v is larger than N_h, denote $N = \min\{N_v, N_h\}$ and assume that U and V represent complete basis respectively of the visible and hidden space.

The joint distribution (2) is conveniently expressed by means of the components of the visible and hidden vectors in these bases:

$$\hat{s}_\alpha = \frac{1}{\sqrt{L}} \sum_{i=1}^{N_v} U_{\alpha i} s_i \qquad \hat{\eta}_\alpha = \frac{1}{\sqrt{L}} \sum_{i=1}^{N_v} U_{\alpha i} \eta_i$$

for $\alpha \in \{1, \dots N_v\}$ and

$$\hat{\sigma}_\alpha = \frac{1}{\sqrt{L}} \sum_{j=1}^{N_h} V_{\alpha j} \sigma_j \qquad \hat{\theta}_\alpha = \frac{1}{\sqrt{L}} \sum_{j=1}^{N_h} V_{\alpha j} \theta_j$$

for $\alpha \in \{1, \dots N_h\}$. These obey the following normalization rules:

$$\sum_{i=1}^{N_v} s_i^2 = L \sum_{\alpha=1}^{N_v} \hat{s}_\alpha^2 \qquad \sum_{j=1}^{N_h} \sigma_j^2 = L \sum_{\alpha=1}^{N_h} \hat{\sigma}_\alpha^2 \qquad \sum_{i=1}^{N_v} \eta_i s_i = L \sum_{\alpha=1}^{N_v} \hat{\eta}_\alpha \hat{s}_\alpha \qquad \sum_{j=1}^{N_h} \theta_i \sigma_i = L \sum_{\alpha=1}^{N_h} \hat{\theta}_\alpha \hat{\sigma}_\alpha$$

we obtain

$$p(\hat{\mathbf{s}}, \hat{\boldsymbol{\sigma}}) = \frac{1}{Z} \exp\left(L \sum_{\alpha=1}^{N} \Big[w_\alpha \hat{s}_\alpha \hat{\sigma}_\alpha + \hat{\eta}_\alpha \hat{s}_\alpha + \hat{\theta}_\alpha \hat{\sigma}_\alpha \Big] - L \sum_{\alpha=1}^{N_v} \frac{\hat{s}_\alpha^2}{2} \right) \delta\Big(L \sum_{\alpha=1}^{N_h} \hat{\sigma}_\alpha^2 - \bar{\sigma}^2 L \Big).$$

From these transformations, we expect $\hat{s}_\alpha, \hat{\sigma}_\alpha$ and also $\hat{\theta}_\alpha, \hat{\eta}_\alpha$ to scale like $\sim L^{-0.5}$. In this representation the SVD modes are coupled by the spherical constraint. To get the distribution of the visible variables alone, we have to integrate over the hidden variables which can be done first by using the Fourier representation of the δ function

$$p(\hat{\boldsymbol{s}}, \hat{\boldsymbol{\sigma}}) = \frac{1}{2i\pi Z} \int_{a-i\infty}^{a+i\infty} dz \exp\left(L\Big(\sum_{\alpha} \Big[w_\alpha \hat{s}_\alpha \hat{\sigma}_\alpha + \hat{\eta}_\alpha \hat{s}_\alpha + \hat{\theta}_\alpha \hat{\sigma}_\alpha \Big] - \sum_{\alpha} \frac{\hat{s}_\alpha^2}{2} - z\big(\sum_{\alpha} \hat{\sigma}_\alpha^2 - \bar{\sigma}^2 \big) \Big) \right), \tag{3}$$

with $a > 0$. With the change of variable $z' = 2\bar{\sigma} z / \Sigma(\hat{\boldsymbol{s}})$ we get

$$p(\hat{\boldsymbol{s}}) = \frac{1}{2iLZ} \Big(\frac{2\pi}{L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}})} \Big)^{N_h/2-1} \exp\left(L \sum_{\alpha=1}^{N_v} \big(\hat{\eta}_\alpha \hat{s}_\alpha - \frac{\hat{s}_\alpha^2}{2} \big) \right) \int_{a-i\infty}^{a+i\infty} \frac{dz}{z^{N_h/2}} \exp\left(\frac{L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}})}{2} \big(z + \frac{1}{z} \big) \right)$$

with

$$\Sigma^2(\hat{\boldsymbol{s}}) \stackrel{\text{def}}{=} \sum_{\alpha=1}^{N} (w_\alpha \hat{s}_\alpha + \hat{\theta}_\alpha)^2.$$

The integration over z can actually be rewritten as (for $N_h \geq 2$)

$$p(\hat{\boldsymbol{s}}) = \frac{(2\pi)^{N_h/2}}{2LZ} \tilde{I}_{N_h/2-1}\big(L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}}) \big) \exp\left(L \sum_{\alpha=1}^{N_v} \big(\hat{\eta}_\alpha \hat{s}_\alpha - \frac{\hat{s}_\alpha^2}{2} \big) \right), \tag{4}$$

where

$$\tilde{I}_\nu(x) = x^{-\nu} I_\nu(x),$$

with the modified Bessel function

$$I_\nu(x) = \frac{x^\nu}{2} \sum_{k=0}^{+\infty} \frac{\big(\frac{x}{2}\big)^{2k}}{k! \Gamma(\nu + k + 1)}.$$

The partition function is also given by means of a single integral after integrating over visible and hidden variables the form (3) after the change $z' = 2z$:

$$Z = \frac{1}{2i\pi} \int_{a-i\infty}^{a+i\infty} dz e^{L\phi(z)}, \tag{5}$$

where

$$\phi(z) = \frac{\bar{\sigma}^2 z}{2} - \frac{\delta}{2}\log(z) + \frac{h_0^2}{2z} + \frac{1}{2}\sum_{\alpha=1}^{N_v}\eta_\alpha^2 + \frac{1}{2}\sum_{\alpha=1}^{N}\Big[\frac{h_\alpha^2}{z-w_\alpha^2} - \frac{1}{L}\log(z-w_\alpha^2)\Big], \tag{6}$$

up to a constant and

$$\begin{aligned}
\delta &\stackrel{\text{def}}{=} (\kappa - \kappa^{-1})\mathbb{1}_{\{N_h > N_v\}},\\
h_0^2 &\stackrel{\text{def}}{=} \mathbb{1}_{\{N_h > N_v\}}\sum_{\alpha=N_v+1}^{N_h}\hat{\theta}_\alpha^2,\\
h_\alpha &\stackrel{\text{def}}{=} \hat{\eta}_\alpha w_\alpha + \hat{\theta}_\alpha.
\end{aligned}$$

2.2 Learning algorithm

The objective of the standard learning procedure of the RBM is to find the set of parameters $\{W, \boldsymbol{\eta}, \boldsymbol{\theta}\}$ such that the likelihood of a given dataset $\underline{s}$ be maximal. This is done by conventional gradient ascent of the log likelihood (LL). The conventional gradient of the LL w.r.t. the parameters is given by

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_{ij}} &= \langle s_i \sigma_j p(\boldsymbol{\sigma}|\boldsymbol{s})\rangle_{\text{Data}} - \langle s_i\sigma_j\rangle_{\text{RBM}}\\
\frac{\partial \mathcal{L}}{\partial \eta_i} &= \langle s_i\rangle_{\text{Data}} - \langle s_i\rangle_{\text{RBM}}\\
\frac{\partial \mathcal{L}}{\partial \theta_j} &= \langle \sigma_j p(\boldsymbol{\sigma}|\boldsymbol{s})\rangle_{\text{Data}} - \langle \sigma_j\rangle_{\text{RBM}}
\end{aligned}$$

This requires to compute various response functions of the RBM and the conditional probability $p(\boldsymbol{\sigma}|\boldsymbol{s})$. As shown in [38, 39] it is convenient to rewrite the gradient in the frame defined by the SVD modes of the weight matrix. As already seen it is here especially adapted since the RBM measure is naturally expressed in this frame. In addition, the specificity of the Gaussian-spherical model is that the joint distribution of the visible variables (4) is invariant w.r.t. a rotation of the singular vector V. This means that we can use a more economical gradient. In addition to modifications of $\{w_\alpha, \hat{\eta}_\alpha, \hat{\theta}_\alpha\}$ we are led to consider infinitesimal rotation $\Omega_{\alpha\beta}$ between modes α and β of the visible SVD basis only. Here $\Omega_{\alpha\beta}$ is a skew-symmetric operator corresponding to the change

$$\begin{aligned}
du_\alpha &= \Omega_{\alpha\beta}u_\beta,\\
du_\beta &= -\Omega_{\alpha\beta}u_\alpha.
\end{aligned}$$

Our simplified LL gradient now reads:

$$\frac{1}{L}\frac{\partial \mathcal{L}}{\partial w_\alpha} = \langle \hat{s}_\alpha \hat{\sigma}_\alpha p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} - \langle \hat{s}_\alpha \hat{\sigma}_\alpha \rangle_{\text{RBM}}$$

$$\frac{1}{L}\frac{\partial \mathcal{L}}{\partial \hat{\eta}_\alpha} = \langle \hat{s}_\alpha \rangle_{\text{Data}} - \langle \hat{s}_\alpha \rangle_{\text{RBM}}$$

$$\frac{1}{L}\frac{\partial \mathcal{L}}{\partial \hat{\theta}_\alpha} = \langle \hat{\sigma}_\alpha p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} - \langle \hat{\sigma}_\alpha \rangle_{\text{RBM}}$$

$$\frac{1}{L}\frac{\partial \mathcal{L}}{\partial \Omega_{\alpha\beta}} = \langle (w_\alpha \hat{s}_\alpha \hat{\sigma}_\beta - w_\beta \hat{s}_\beta \hat{\sigma}_\alpha) p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} - \langle (w_\alpha \hat{s}_\alpha \hat{\sigma}_\beta - w_\beta \hat{s}_\beta \hat{\sigma}_\alpha)\rangle_{\text{RBM}}$$

with

$$p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}}) = \frac{\bar{\sigma}^2 L}{(2\pi)^{N_h/2}} \frac{\exp\Big(L\sum_\alpha w_\alpha \hat{s}_\alpha \hat{\sigma}_\alpha + \hat{\theta}_\alpha \hat{\sigma}_\alpha\Big)}{\tilde{I}_{N_h/2-1}\big(L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}})\big)} \delta\Big(L\sum_\alpha \hat{\sigma}_\alpha^2 - \bar{\sigma}^2 L\Big).$$

This results in the following (continuous) update equations for the parameters and the dataset $\mathcal{S}$:

$$\frac{dw_\alpha}{dt} = \langle \hat{s}_\alpha \hat{\sigma}_\alpha p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} - \langle \hat{s}_\alpha \hat{\sigma}_\alpha \rangle_{\text{RBM}}, \tag{7}$$

$$\frac{d\hat{\eta}_\alpha}{dt} = \langle \hat{s}_\alpha \rangle_{\text{Data}} - \langle \hat{s}_\alpha \rangle_{\text{RBM}} - \sum_\beta \Omega_{\alpha\beta}\hat{\eta}_\beta, \tag{8}$$

$$\frac{d\hat{\theta}_\alpha}{dt} = \langle \hat{\sigma}_\alpha p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} - \langle \hat{\sigma}_\alpha \rangle_{\text{RBM}}, \tag{9}$$

$$\frac{d\hat{s}_\alpha^k}{dt} = -\sum_\beta \Omega_{\alpha\beta}\hat{s}_\beta^k, \qquad \forall k \in \mathcal{S}, \tag{10}$$

with

$$\Omega_{\alpha\beta} = \langle (w_\alpha \hat{s}_\alpha \hat{\sigma}_\beta - w_\beta \hat{s}_\beta \hat{\sigma}_\alpha) p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} - \langle (w_\alpha \hat{s}_\alpha \hat{\sigma}_\beta - w_\beta \hat{s}_\beta \hat{\sigma}_\alpha)\rangle_{\text{RBM}} \tag{11}$$

and where dt/L represents the learning rate. Here the last update equation corresponds to simply adapting the data projection on the rotated basis. Note that the same is done also for the second update equation concerning the field projection $\hat{\eta}$, which is optional here but coherent with the conventional update rules and useful in practice. Note also that the singularity of the conventional gradient observed in [38] for pairs of modes with identical singular values has disappeared. Hence, computing the gradient requires to evaluate one and two-points correlation functions of SVD variables.

As seen in the previous section, the LL takes the form

$$\mathcal{L} = \Big\langle \log\Big(\tilde{I}_{N_h/2-1}\Big[L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}})\Big]\Big) + L\sum_\alpha \big(\hat{\eta}_\alpha \hat{s}_\alpha - \frac{\hat{s}_\alpha^2}{2}\big)\Big\rangle_{\text{Data}} - \log\big(Z\big).$$

Compared to the simple Gaussian RBM likelihood, we see one important difference: eigenvalues of W do interact, in particular via the empirical term which is now a nonlinear, monotonically increasing function of $\Sigma(\hat{\boldsymbol{s}})$. With help of the identity

$$\frac{d\tilde{I}_\nu(x)}{dx} = x\tilde{I}_{\nu+1}(x),$$

we get from this the gradient of the log likelihood in the form:

$$\frac{\partial \mathcal{L}}{\partial w_a} = \bar{\sigma}\Big\langle \hat{s}_\alpha(w_\alpha \hat{s}_\alpha + \hat{\theta}_\alpha)\frac{I_{N_h/2}\big(L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}})\big)}{\Sigma(\hat{\boldsymbol{s}})I_{N_h/2-1}\big(L\bar{\sigma}\Sigma(\hat{\boldsymbol{s}})\big)}\Big\rangle_{\text{Data}} - \frac{\partial \log(Z)}{\partial w_\alpha}$$

Using the following asymptotic expression for large ν (see e.g. [42])

$$I_\nu(\nu z) \sim \frac{1}{\sqrt{2\pi\nu}}\frac{e^{\nu\eta}}{(1+z^2)^{1/4}} \qquad \text{with} \qquad \eta = \sqrt{1+z^2} + \log\frac{z}{1+\sqrt{1+z^2}},$$

resulting from a saddle point approximation of the modified Bessel function, we obtain the asymptotic expression

$$\langle \hat{s}_\alpha \hat{\sigma}_\beta p(\hat{\boldsymbol{\sigma}}|\hat{\boldsymbol{s}})\rangle_{\text{Data}} = \bar{\sigma}\left\langle \hat{s}_\alpha \frac{w_\beta \hat{s}_\beta + \hat{\theta}_\beta}{1+\sqrt{1+\bar{\sigma}^2\Sigma(\hat{\boldsymbol{s}})^2}}\right\rangle_{\text{Data}},$$

valid for large L.

The remaining point to address now in order to be able to train such a machine is the estimation of the partition function and its derivatives. For the rest of the paper, the local fields $\hat{\eta}_\alpha$ on the visible variables will be set to zero to lighten the presentation.

3 Thermodynamical properties

The expression of the partition function given by eq. (5-6) indicates that the physical properties of the Gaussian-spherical RBM depend only on the spectrum of its weight matrix in absence of the fields as for the ordinary spherical model [43]. Standard treatments of the spherical model (see [40, 44, 45]) rely on a saddle point approximation of the contour integral representation of Z given by eq. (5). Here we recall and straightforwardly adapt these arguments to our needs by making simple assumptions on the limit spectrum of W when $L \to \infty$. Note also that variational properties of the free energy of bipartite models with spherical constraints in thermodynamic limit have been established in the recent years as in [46] for (p, q)-spin bipartite at high temperature or in [47, 48] for the RBM with two spherical constraints in terms of the spectral density of the coupling matrix. The Gaussian-spherical setting that we consider is easier to analyze since only one complex integral instead of two for the spherical-spherical RBM as detailed in [47], is necessary to express the partition function as we will see below. In a second step this will lead us to establish incidentally a connections with condensation phase transition analyzed in the context of factorized steady states [41].

3.1 Ferromagnetic transition

First notice that ϕ given in (6) is convex on the domain of interest:

$$\phi''(z) = \frac{\delta}{2z^3} + \frac{h_0^2}{z^3} + \sum_\alpha \Big[\frac{h_\alpha^2}{(z-w_\alpha^2)^3} + \frac{1}{2L}\frac{1}{(z-w_\alpha^2)^2} \Big] > 0, \qquad \text{for } z > w_{\max}^2,$$

with $w_{\max}$ the highest singular value, so there is only one solution z_0 to the saddle point equation allowing for the following approximation:

$$Z \sim_{L\to\infty} \frac{\exp\big(L\phi(z_0)\big)}{\sqrt{2\pi L|\phi''(z_0)|}}.$$

At the saddle point the free energy per degree of freedom is given by

$$f = -\phi(z_0, \eta, \theta).$$

From these quantities we can in principle get all the needed response functions (see Appendix A). We will focus here on the computation of the spontaneous magnetizations as a function of the spectrum of W. Their expressions can be obtain using

$$\langle \hat{s}_\alpha \rangle = -\frac{\partial f}{\partial \hat{\eta}_\alpha} = w_\alpha \frac{h_\alpha}{z_0 - w_\alpha^2}, \tag{12}$$

$$\langle \hat{\sigma}_\alpha \rangle = -\frac{\partial f}{\partial \hat{\theta}_\alpha} = \frac{h_\alpha}{z_0 - w_\alpha^2}. \tag{13}$$

This gives us relations between magnetizations and z_0. In order to analyze further the thermodynamic properties of the system some assumptions have to be made on the spectral properties of W. Let us define the spectral density (SD) associated to WW^T:

$$\rho_L(E) \stackrel{\text{def}}{=} \frac{1}{L}\sum_{\alpha=1}^{N_h} \delta\big(E - w_\alpha^2\big),$$

(which includes zero modes $w_\alpha = 0$ for $\alpha > N_v$ whenever $N_h > N_v$). In the thermodynamic limit it is assumed that the SD tends to a well defined limit distribution

$$\lim_{L\to\infty} \rho_L(E) = \rho(E).$$

This leads to

$$\phi(z) = \frac{1}{2}\bar{\sigma}^2 z + \frac{1}{2}\int_0^{E_{\max}} dE\rho(E)\Big(\frac{h(E)^2}{z-E} - \log(z-E)\Big),$$

with some upper bound $E_{\max}$ of the SD, and where $h(E)$ is any smooth function taking the value $\sqrt{L}h_\alpha$ for $E = w_\alpha^2$ at finite L, h_α being expected to be $O(1/\sqrt{L})$ in general. $\rho(E)h(E)^2$ represents the SD of the external field. As in [44] for instance, we have to distinguish between a situation where the SD has isolated dominant modes and the situation with just

a continuous bulk of modes bounded by $E_{\max}$. When the SD has no isolated dominant mode we look for a solution z_0 to the saddle point equation

$$\begin{aligned}\phi'(z) &= \frac{1}{2}\big(g_1(z) - g_2(z)\big) \\ &= 0\end{aligned}$$

where

$$\begin{aligned}g_1(z) &\stackrel{\text{def}}{=} \bar{\sigma}^2 - \int_0^{E_{\max}} dE \frac{\rho(E)}{(z-E)^2} h(E)^2, \\ g_2(z) &\stackrel{\text{def}}{=} \int_0^{E_{\max}} dE \frac{\rho(E)}{z-E},\end{aligned}$$

Let us call

$$\bar{\sigma}_c^2 \stackrel{\text{def}}{=} g_2(E_{\max}) \tag{14}$$

The properties of the system depends on the behavior of $\rho(E)$ near $E_{\max}$. A thorough discussion of its influence on the physical properties of the spherical model can be found in [43]. Here we restrict the discussion to behavior of the type $\rho(E) \sim (E_{\max} - E)^\gamma$ with the exponent $\gamma > -1$. This cover various study cases like for instance d-dimensional regular lattices $\gamma = d/4 - 1$ or $\gamma = 1/2$ for i.i.d. random matrices. In order to get closed form expressions we shall consider the following beta distributions for the SD:

$$\rho(E) = \frac{\kappa}{B(1-\gamma, \gamma+1)} \frac{E^{-\gamma}(E_{\max} - E)^\gamma}{E_{\max}} \qquad \text{with} \qquad \gamma \in]-1, 1[, \tag{15}$$

$$\rho(E)h(E)^2 = \frac{h^2}{B(\beta+1, 1-\beta)} \frac{E^{\beta}(E_{\max} - E)^{-\beta}}{E_{\max}} \qquad \text{with} \qquad \beta \in]-1, 1[, \tag{16}$$

where the beta function takes here the special form

$$B(1-\gamma, 1+\gamma) = \frac{\gamma\pi}{\sin(\gamma\pi)},$$

and with h^2 the squared norm of the external field. $\rho(E)h(E)^2$ represents the SD of the external fields. This setting can be useful to study the response function at the top of the spectrum, by simply letting $\beta \to 1$. $\bar{\sigma}_c$ is infinite for $\gamma \leq 0$ and finite otherwise with

$$\bar{\sigma}_c^2 = \frac{1}{E_{\max}}\Big(\delta + \frac{\kappa}{\gamma}\Big), \qquad \gamma > 0$$

in the latter case. The different scenarios for obtaining z_0 are sketched on Figure 2.

- When $\gamma \leq 0$, g_2 diverges close to $E_{\max}$ and therefore the intersection A with g_1 always converges to the point $B = (z_0 > E_{\max}, \bar{\sigma}^2)$. In that case, there is no way that $(z - E)$ goes to zero for $E \leq E_{\max}$ when $h \to 0$ and therefore all the magnetizations (12,13) vanish.

- When $\gamma > 0$ we have to distinguish between two cases. First, we consider condensation on modes that have $E < E_{\max}$ by applying small vanishing fields on these modes. In that case, since when $h \to 0$ we have $z_0 \geq E_{\max}$, again the magnetization will simply vanish since the denominator in (12,13) is finite and non-zero. For modes at $E_{\max}$ if $\bar{\sigma} < \bar{\sigma}_c$, then $z_0 \to E > E_{\max}$ when we put the field to zero therefore giving the same results as in the first scenario. Now, if instead $\bar{\sigma} \geq \bar{\sigma}_c$, $z_0 \to E_{\max}$ $h(E_{\max})\rho(E_{\max})/(z_0 - E_{\max})$ has a finite limits given below. We obtain a spontaneous magnetization in that case.

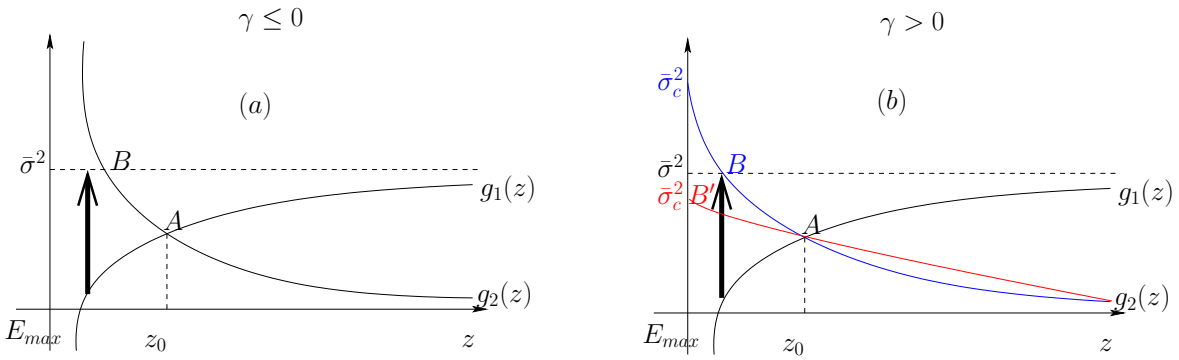

Figure 2. Various scenarios for the saddle points solution z_0. Point A corresponds to the intersection between g_1 and g_2 at finite field $h(E)$, the black arrow the movement of g_1 when the fields tend to zero while points B and B' corresponds to the limit intersection at vanishing fields when respectively $\bar{\sigma} < \bar{\sigma}_c$ (in blue) and $\bar{\sigma} > \bar{\sigma}_c$ (in red).

The spontaneous magnetization is obtained as follows for the cases where a saddle point solution exists. If a field $h_\alpha = h_{\max}$ is concentrated on the largest mode, this is equivalent to choose $h(E)$ such that $\rho(E)h(E)^2 = h_{\max}^2 \delta(E - E_{\max})$. We get from the saddle point equation:

$$\bar{\sigma}^2 - \frac{h_{\max}^2}{(z_0(h_{\max}) - E_{\max})^2} = \bar{\sigma}_c^2.$$

Now, when $z_0(h_{\max}) \to E_{\max}$ and we let $h_{\max} \to 0$ we obtain

$$\lim_{h_{\max} \to 0} \frac{h_{\max}}{z_0(h_{\max}) - E_{\max}} = \sqrt{\bar{\sigma}^2 - \bar{\sigma}_c^2}.$$

Eliminating $z_0 - E$ in (12,13) yields the spontaneous magnetization

$$\langle \hat{s}_\alpha \rangle = w_{\max} \bar{\sigma} \sqrt{\bar{\sigma}^2 - \bar{\sigma}_c^2},$$

$$\langle \hat{\sigma}_\alpha \rangle = \sqrt{\bar{\sigma}^2 - \bar{\sigma}_c^2},$$

and

$$\langle \hat{s}_\alpha \hat{\sigma}_\beta \rangle = \delta_{\alpha\beta} w_\alpha (\bar{\sigma}^2 - \bar{\sigma}_c^2).$$

When the highest mode acquires a macroscopic magnetization in the ferromagnetic phase, the resulting distribution $p(\boldsymbol{s})$ becomes bimodal along this direction. It is also worth noticing

that when the highest mode is degenerated n times, in absence of any external fields the system has an $O(n)$ symmetry corresponding to rotations in the subspace defined by these vectors. This results in that case into a distribution concentrated on a n-dimensional sphere in the ferromagnetic phase. The specific shape of the condensed distribution will be studied in the next subsection.

3.2 Condensation mechanism in thermodynamic limits

The scaling form (15,16) of the SD of singular values and field densities allows us to make an explicit connection with scaling function derived in a different context, namely condensation of factorized steady states [41]. After dropping irrelevant terms and making the change $z' = (z - E_{\max})/E_{\max}$ while absorbing $E_{\max}$ in the definition of $\bar{\sigma}$ we get the following form of the partition function:

$$Z_{L,N}[\bar{\sigma}, h] = \frac{1}{2i\pi}\int_{-i\infty}^{i\infty} dz e^{\frac{L}{2}\phi(z,\bar{\sigma},h)},$$

with (see Appendix B)

$$\phi(z,\bar{\sigma},h) = \bar{\sigma}^2(z+1) - \frac{\kappa}{\gamma}\int_0^z du\Big[1-\Big(\frac{u}{1+u}\Big)^{\gamma}\Big] + \frac{h^2}{\beta}\Big[\Big(\frac{1+z}{z}\Big)^{\beta} - 1\Big]. \tag{17}$$

For large L the rescaling $L^{1/(1+\gamma)}z \to z$ leads to the scaling behavior

$$Z_{L,N}(\bar{\sigma},h) = e^{\frac{L}{2}\left(\bar{\sigma}^2 - \frac{h^2}{\beta}\right)}\left(L^{-\frac{1}{\gamma+1}}V_{\gamma,\beta}\Big(L^{\frac{\gamma}{1+\gamma}}(\bar{\sigma}^2-\bar{\sigma}_c^2), L^{\frac{2+\beta+\gamma}{1+\gamma}}h^2\Big) + \mathcal{O}\Big(\frac{1}{L^{\frac{2}{1+\gamma}}}\Big)\right),$$

valid for $(\gamma,\beta) \in (]-1,0[\cup]0,1[)^2$, where the following scaling function has been introduced

$$\begin{aligned} V_{\gamma,\beta}(x,y) &= \int_{-i\infty}^{i\infty} dz \exp\Big(\frac{1}{2}xz + bz^{\gamma+1} + c\frac{y}{z^{1+\beta}}\Big), \\ &= \frac{1}{\pi}\int_0^{\infty} du\; e^{-b_2u^{\gamma+1} - c_2\frac{y}{u^{\beta+1}}}\cos\Big(\frac{xu}{2} - b_1u^{\gamma+1} + c_1\frac{y}{u^{\beta+1}}\Big), \end{aligned}$$

with

$$\begin{aligned} b &= \frac{\kappa}{2\gamma(\gamma+1)} \qquad & b_1 &= b\cos\big(\frac{\gamma\pi}{2}\big) \qquad & b_2 &= b\sin\big(\frac{\gamma\pi}{2}\big) \\ c &= \frac{1}{2\beta} \qquad & c_1 &= c\cos\big(\frac{\beta\pi}{2}\big) \qquad & c_2 &= c\sin\big(\frac{\beta\pi}{2}\big) \end{aligned}$$

and where the change of variable $z = \pm iu$ is used for $Im(z) \pm 0$. In [41] the same scaling function (at $y = 0$) is encountered albeit in a different context. We can therefore closely follow their analysis to describe the transition to ferromagnetic order of the present spherical model. As seen previously, when $\gamma > 0$ there is a possibility for dominant modes to generate

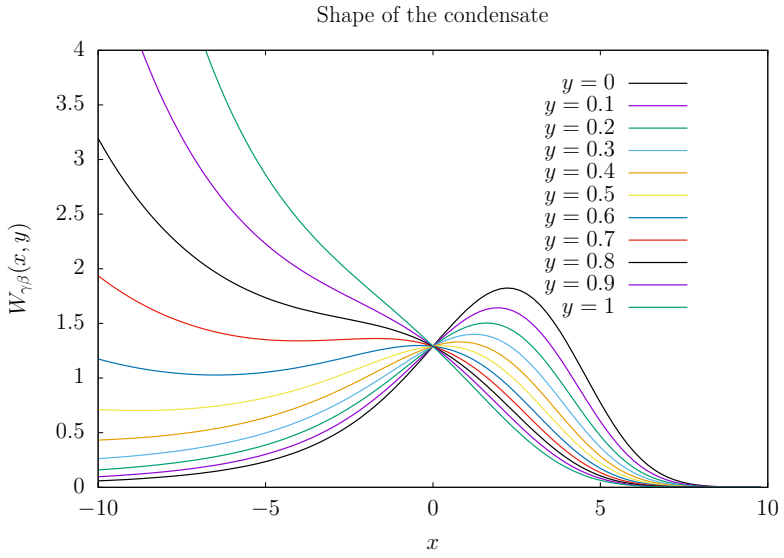

Figure 3. Shape of the condensate distribution ($\gamma = 0.5$) along a given mode α as the scaled distance $y = L^{\frac{1}{1+\gamma}}(1-\epsilon_\alpha)$ from the upper boundary is increased, x being the scaled fraction of variance $L^{\frac{\gamma}{1+\gamma}}(\sigma_\alpha^2 - V_{\rm ex})$ along this mode. The bump in the distribution disappear for $y \geq y_{0.5} \simeq 0.72$, while the second mode correspond to the range $y \in [1.77, 2.81[$ for $\gamma = 0.5$.

ferromagnetic order when $\bar{\sigma}_c < \bar{\sigma}$. In presence of an external field ($h^2 > 0$) there is always a solution to the saddle point equation and no transition occurs at $\bar{\sigma}_c$. Instead, in absence of external fields and $\gamma > 0$, there is no solution to the saddle point equation when $\bar{\sigma}_c < \bar{\sigma}$ while there is always one in the opposite case, and the transition corresponds to the onset of ferromagnetic order materialized by condensation along the dominant modes. In that case a finite fraction of the overall variance of the distribution is captured by one or possibly a small number of modes.

In order to study the condensate we need to express the marginal probabilities $p_{\hat{s}_\alpha}(x) \stackrel{\text{def}}{=} P(\hat{s}_\alpha = x)$ and $p_{\hat{\sigma}_\alpha}(x) \stackrel{\text{def}}{=} P(\hat{\sigma}_\alpha = x)$. For any given mode α we have (in absence of external fields)

$$p_{\hat{s}_\alpha}(x) = \int dy p_{\hat{\sigma}_\alpha}(y) e^{L\left(\sqrt{\epsilon_\alpha} xy - \frac{x^2}{2}\right)}$$

$$p_{\hat{\sigma}_\alpha}(x) = \frac{Z_{L,N-1}\left(\bar{\sigma}^2 - x^2\right)}{Z_{L,N}} \exp\left(\frac{L}{2}\epsilon_\alpha x^2\right),$$

with $\epsilon_\alpha = E_\alpha / E_{\max}$ and

$$Z_{L,N} = \int_0^\infty dx Z_{L,N-1}\left(\bar{\sigma}^2 - x^2\right) \exp\left(\frac{L}{2}\epsilon_\alpha x^2\right),$$

where it is assumed that $Z_{L,N}$ corresponds to the system with one single mode at ϵ_α added

to the SD (15), while $Z_{L,N-1}$ corresponds to the SD (15) alone. Let us call

$$V_{\rm ex} = \bar{\sigma}^2 - \bar{\sigma}_c^2,$$

the "excess of variance" in the system. We get for the condensate the following behavior

$$p_{\hat{\sigma}_\alpha}(x) \propto W_\gamma\Big(L^{\frac{\gamma}{1+\gamma}}(x^2 - V_{\rm ex}), L^{\frac{1}{1+\gamma}}(1-\epsilon_\alpha)\Big)$$

with now

$$W_\gamma(x,y) = e^{\frac{-xy}{2}} \int_0^\infty du\ e^{-b_2 u^{\gamma+1}} \cos\Big(\frac{xu}{2} - b_1 u^{\gamma+1}\Big)$$

So $W_\gamma(x,y) = e^{-xy/2} V_\gamma(-x/2)$ whose plot is shown on Figure 3 and which asymptotic behavior for large x is given in Appendix B. This help us to determine how many modes condense and the shape of the distribution along these modes, in the vincinity of the upper boundary of the spectrum corresponding to $\epsilon_\alpha = 1$. Strictly speaking the bump observed on Figure 3 represents the condensation of a mode only for $y = 0$, because as soon as y is strictly positive $W_\gamma(x,y) \sim_{x\to-\infty} e^{-xy/2}/|x|^{\gamma+2}$, which means that the contribution of the bump to the distribution is suppressed exponentially by a factor $\exp\big(-L^{\frac{1}{\gamma+1}} V_{\rm ex} y/2\big)$ by comparison to contributions near $\sigma_\alpha^2 = 0$. Still we see that the bump is present for some values $y \in [0, y_\gamma[$. To know to which modes this corresponds to, first note that $y = L^{\frac{1}{1+\gamma}}(1-\epsilon)$ is actually a measure of the rank from the top of the spectrum. Given the SD (15) the kth mode is actually located in the range $y \in [y_\gamma^{(k)}, y_\gamma^{(k+1)}[$ with

$$y_k \stackrel{\rm def}{=} \Big[k \frac{\gamma(\gamma+1)\pi}{\sin(\gamma\pi)}\Big]^{\frac{1}{\gamma+1}},$$

corresponding to a value of ϵ s.t. $L\int_\epsilon^1 du \rho(u) = k$ for large L and finite k. It can be checked numerically that y_γ is always below $y_\gamma^{(1)}$ for $\gamma \in]0,1[$, which means that the bump concerns only the highest mode.

For modes which are detached above the bulk we have to consider the situation with $y = L^{\frac{1}{1+\gamma}}(1-\epsilon) < 0$. In that case $p_{\hat{\sigma}_\alpha}(x)$ has a Bell shape centered around $x^2 = V_{\rm ex}$, and according to the asymptotic behavior of V_γ given in appendix, $p_{\hat{\sigma}_\alpha}(x)$ decays like $1/|V_{\rm ex} - x^2|^{\gamma+2}$ when $x \to 0$ and like $\exp\Big(-c_2 L (x^2 - V_{\rm ex})^{\frac{\gamma+1}{\gamma}} + \frac{L}{2}(x^2 - V_{\rm ex})(\epsilon - 1)\Big)$ for $x^2 \gg V_{\rm ex}$ ($(\gamma+1)/\gamma$ is always greater that 1 for $\gamma \in]0,1[$).

4 Doubly degenerate spherical RBM

Instead of using the saddle point approximation we remark that closed-form expressions of the partition function and of the LL can be obtained in the case where the spectrum of the weight matrix has discrete levels with a multiplicity of 2 per level. Defining an RBM obeying this constraint can be done simply by a duplication of both the input and hidden layer with two identical blocks of the weight matrix but keeping one single spherical constraint on the hidden layer.

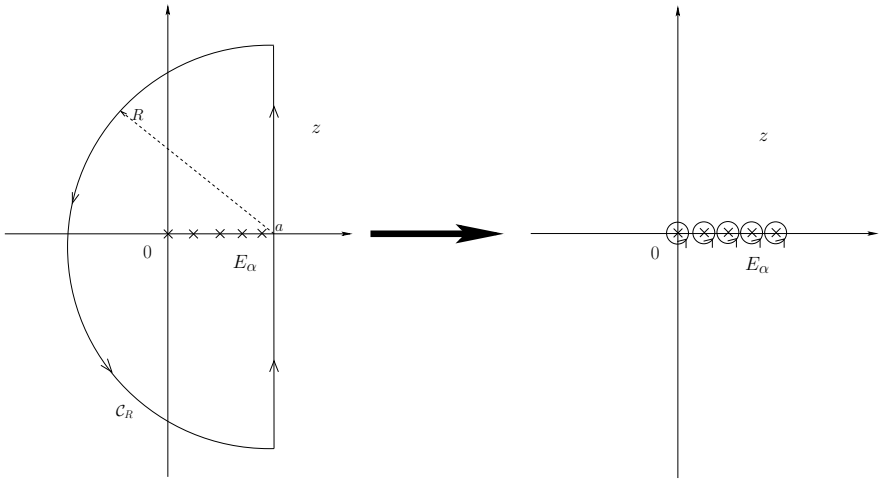

Figure 4. Integration contour deformation.

4.1 Dual formulation

Let $K = \max(0, (N_h - N_v)/2)$ and let now $N = \min(N_v, N_h)/2$ represent half the rank of W. The weight matrix takes now the form:

$$W = \sum_{\substack{\alpha=1 \\ \omega\in\{1,2\}}}^{N} w_\alpha u^{\alpha,\omega} v^{\alpha,\omega}.$$

Given this we then have the following form for the partition function:

$$Z = \frac{1}{2i\pi}\int_{a-i\infty}^{a+i\infty} \frac{dz}{z^K} \prod_{\alpha=1}^{N} (z - E_\alpha)^{-1} e^{L\phi(z)}$$

with now

$$\phi(z) = \frac{\bar{\sigma}^2 z}{2} + \frac{h_0^2}{2z} + \frac{1}{2}\sum_{\alpha=1}^{N} \frac{h_\alpha^2}{z - E_\alpha},$$

with

$$h_\alpha^2 = h_{\alpha,1}^2 + h_{\alpha,2}^2$$

and

$$h_{\alpha,\omega} = w_\alpha \eta_{\alpha,\omega} + \theta_{\alpha,\omega}, \qquad \alpha = 1, \dots N$$

$$h_0^2 = \sum_{\alpha=N+1}^{N+K} \left(\theta_{\alpha,1}^2 + \theta_{\alpha,2}^2\right).$$

To evaluate Z we can deform the integration contour to the half circle $\mathcal{C}_R$ as shown on Figure 4:

$$Z = \lim_{R\to\infty} \frac{1}{2i\pi} \oint_{\mathcal{C}_R} \frac{dz}{z^K} \prod_{\alpha} (z - E_\alpha)^{-1} e^{L\phi(z)},$$

thanks to the following bound for the contribution on the half circle for R sufficiently large:

$$R\left|z^{-K}\prod_{\alpha}(z-E_\alpha)^{-1}e^{L\phi(z)}\right| \leq \frac{A}{R^{N_h-1}} \to_{R\to\infty} 0.$$

Then, since the integrand is holomorphic everywhere inside the domain enclosed by $\mathcal{C}_R$ except on the singularities $z = E_\alpha$, we can deform the contour as shown on Figure 4 in terms of small anti-clockwise circles $\mathcal{C}_{\epsilon,\alpha}$ of radius ϵ around each singularities including $z = 0$ for $\mathcal{C}_{\epsilon,0}$, ϵ being small enough such that each $\mathcal{C}_{\epsilon,\alpha}$ encloses one single singularity. Z is then expressed as

$$Z = \frac{1}{2i\pi}\sum_{\alpha=0}^{N}\oint_{\mathcal{C}_{\alpha,\epsilon}}\frac{dz}{z^K}\prod_{\beta}(z-E_\beta)^{-1}e^{L\phi(z)},$$

which after expanding for each contour the enclosed singular part in the exponential reads

$$Z = \sum_{n=0}^{\infty}\Big(\frac{\big(Lh_0^2\big)^n}{4^n n!^2}f_0^{(n+K-1)}(0) + \sum_{\alpha=1}^{N}\frac{\big(Lh_\alpha^2\big)^n}{4^n n!^2}f_\alpha^{(n)}(0)e^{\frac{L}{2}\bar{\sigma}^2E_\alpha}\Big),$$

where

$$f_0(z) = \prod_{\beta}\frac{1}{z-E_\beta}\exp\Big(\frac{L}{2}\Big[\bar{\sigma}^2 z + \sum_{\beta}\frac{h_\beta^2}{z-E_\beta}\Big]\Big),$$

$$f_\alpha(z) = \frac{1}{(z-E_\alpha)^K}\prod_{\beta\neq\alpha}\frac{1}{z+E_\alpha-E_\beta}\exp\Big(\frac{L}{2}\Big[\bar{\sigma}^2 z + \frac{h_0^2}{z+E_\alpha} + \sum_{\beta\neq\alpha}\frac{h_\beta^2}{z+E_\alpha-E_\beta}\Big]\Big).$$

It is rather tedious to write down this expression for Z more explicitly, instead at this point let us consider the case when all $h_\alpha = 0$. Then we simply get

$$Z = \sum_{\alpha=1}^{N}\frac{1}{E_\alpha^K}\Big(\exp\big(\frac{\bar{\sigma}^2 LE_\alpha}{2}\big) - \sum_{k=0}^{K-1}\frac{1}{k!}\big(\frac{\bar{\sigma}^2 LE_\alpha}{2}\big)^k\Big)\prod_{\beta\neq\alpha}(E_\alpha-E_\beta)^{-1}. \tag{18}$$

Reducing this expression to the same denominator leads us to express the partition function as a ratio of two determinants:

$$Z = \frac{\begin{pmatrix} 1 & E_1 & E_1^2 & \dots & E_1^{N-2} & \sum_{k=N-1}^{\infty}\frac{1}{(k+K)!}\big(\frac{\bar{\sigma}^2 LE_1}{2}\big)^k \\ 1 & E_2 & E_2^2 & \dots & E_2^{N-2} & \sum_{k=N-1}^{\infty}\frac{1}{(k+K)!}\big(\frac{\bar{\sigma}^2 LE_2}{2}\big)^k \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & E_N & E_N^2 & \dots & E_N^{N-2} & \sum_{k=N-1}^{\infty}\frac{1}{(k+K)!}\big(\frac{\bar{\sigma}^2 LE_N}{2}\big)^k \end{pmatrix}}{\begin{pmatrix} 1 & E_1 & E_1^2 & \dots & E_1^{N-2} & E_1^{N-1} \\ 1 & E_2 & E_2^2 & \dots & E_2^{N-2} & E_2^{N-1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & E_N & E_N^2 & \dots & E_N^{N-2} & E_N^{N-1} \end{pmatrix}}$$

This ratio is actually a weighted sum of particular Schur polynomials (see e.g. [49]), each one being a positive symmetric function of the energy levels E_α. There is a generating function for these thanks the following identity

$$\begin{pmatrix} 1 & E_1 & E_1^2 & \dots & E_1^{N-2} & \frac{1}{1-tE_1} \\ 1 & E_2 & E_2^2 & \dots & E_2^{N-2} & \frac{1}{1-tE_2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & E_N & E_N^2 & \dots & E_N^{N-2} & \frac{1}{1-tE_N} \end{pmatrix} = \prod_{\alpha<\beta}(E_\alpha - E_\beta)\frac{t^{N-1}}{\prod_\alpha(1-tE_\alpha)},$$

we have

$$Z = \sum_{\{n_\alpha\}} \frac{\left(\frac{\bar\sigma^2 L}{2}\right)^{K+N-1}}{(\sum_\alpha n_\alpha + K + N - 1)!} \prod_{\alpha=1}^{N} \left(\frac{L}{2}\bar\sigma^2 E_\alpha\right)^{n_\alpha}, \tag{19}$$

where the n_α's run over $\mathbb{N}$.

4.2 Urn model interpretation

Expression (19) allows us to make the connection with another type of models studied in statistical physics, namely urn models (see e.g. [50]) which generalize the Ehrenfest model to an extensive number of urns. These are in fact a special case of queuing network processes [51] well studied in probability theory. In the queuing theory, general class of queuing networks have been identified which have simple explicit steady state measures [52]. The first one is the so called Jackson network with exponential service rates [53] possibly open or closed. Up to a multiplicative constant, the form (19) coincides with the normalization of the invariant measure of a closed Jackson network of queues. Here each index α refers to a queue characterized by a service rate μ_α and n_α is interpreted as the number of clients in the queue. These queues are assembled into a network by fixing a set of routing probabilities among them. Many different configurations can possibly lead to the same measure corresponding to (19). Let us give an instance of such a network fulfilling this constraint. In addition of the queues already defined we add an additional one with index $\alpha = 0$ corresponding to the reservoir, n_0 being its number of clients supposedly large. The routing is then simply defined by the given ordering of the queues: a client in queue α is routed to queue $(\alpha + 1)\ mod(N)$, which means that the system is closed and the queues are arranged along a circle. Let $N_\infty = n_0 + \sum_{\alpha>0} n_\alpha$ represent the total number of clients, considered arbitrarily large. Consider the following service rates:

$$\mu_\alpha = \begin{cases} E_\alpha^{-1} & \forall \alpha \in \{1, \dots N\}, \\ \dfrac{\bar\sigma^2 L}{2(N_\infty - n_0 + K + N - 1)} & \text{for } \alpha = 0. \end{cases}$$

We are then in the conditions of the Jackson theorem, namely that each service rate depends only the state of the corresponding queue and that there exists a set of actually constant

arrival rates $\lambda_\alpha = \lambda$, satisfying the so-called traffic equation (i.e. flux conservation on the network). Hence the corresponding measure has the following form

$$P(\boldsymbol{n}) = \frac{1}{\tilde{Z}(N_\infty)} \prod_{n=0}^{n_0-1} \frac{2\lambda(N_\infty - n + K + N - 1)}{\bar{\sigma}^2 L} \prod_{\alpha=1}^{N} \Big(\frac{\lambda}{\mu_\alpha}\Big)^{n_\alpha} \delta\Big(N_\infty - \sum_{\alpha=0}^{N} n_\alpha\Big),$$

where λ is arbitrary, with (after re-arranging the summation)

$$\tilde{Z}(N_\infty) = \lambda^{N_\infty} \sum_{\{n_\alpha, \alpha>0\}} \mathbb{1}_{\{\sum_\alpha n_\alpha \leq N_\infty\}} \Big(\frac{L\bar{\sigma}^2}{2}\Big)^{\sum_\alpha n_\alpha - N_\infty} \frac{(N_\infty + K + N - 1)!}{\big(\sum_\alpha n_\alpha + K + N - 1\big)!} \prod_{\alpha=1}^{N} E_\alpha^{n_\alpha}.$$

So if we now let the size of the reservoir (controlled by N_∞) become sufficiently large, we see that up to an irrelevant N_∞-dependent factor, $\tilde{Z}(N_\infty)$ coincides with Z:

$$Z = \lim_{N_\infty \to \infty} \frac{\lambda^{-N_\infty} (\bar{\sigma}^2 L/2)^{N_\infty + N + K - 1}}{(N_\infty + N + K - 1)!}\ \tilde{Z}(N_\infty).$$

From the practical point of view, Z can be computed with arbitrary precision, thanks to the following recursion. Let

$$Z_{l,m}[E_1, \ldots, E_l] \stackrel{\text{def}}{=} \sum_{n_1, \ldots, n_l} \mathbb{1}_{\{\sum_{\alpha=1}^{l} n_\alpha = m\}} \prod_{\alpha=1}^{l} E_\alpha^{n_\alpha}.$$

We have

$$Z = \sum_{m=0}^{\infty} \frac{\big(\frac{\bar{\sigma}^2 L}{2}\big)^{m+K+N-1}}{(m + K + N - 1)!} Z_{N,m}\big[E_1, \ldots, E_N\big]. \tag{20}$$

In order to compute Z numerically we make use of the following recursion:

$$Z_{l+1,m}[E_1, \ldots, E_{l+1}] = \sum_{k=0}^{m} E_{l+1}^k Z_{l,m-k}[E_1, \ldots, E_l].$$

Thanks to this recursion, if now we fix an upper bound $M = O(L)$ (in the condensed phase) of the maximal number of clients in order to reach a given precision for Z, we end up with a complexity $\mathcal{O}(L^2)$ to estimate the partition function. On Figure 5 is shown the finite size dependence of the two-point function $\langle \hat{s}_\alpha \hat{\sigma}_\alpha \rangle$ using these recursions.

4.3 The ferromagnetic transition as a Bose-Einstein condensation

It is known for a very long time that the spherical model is related to the ideal Bose gas and that the transition is analogous to the Bose-Einstein condensation (see [43] and references herein). In this queueing process language we can make it very explicit. In its original formulation, the ferromagnetic transition is associated to sharp increase of the magnetization projected on the dominant mode, which results for the thermal expectation $\langle \hat{s}_\alpha^2 \rangle_{\text{RBM}}$ along

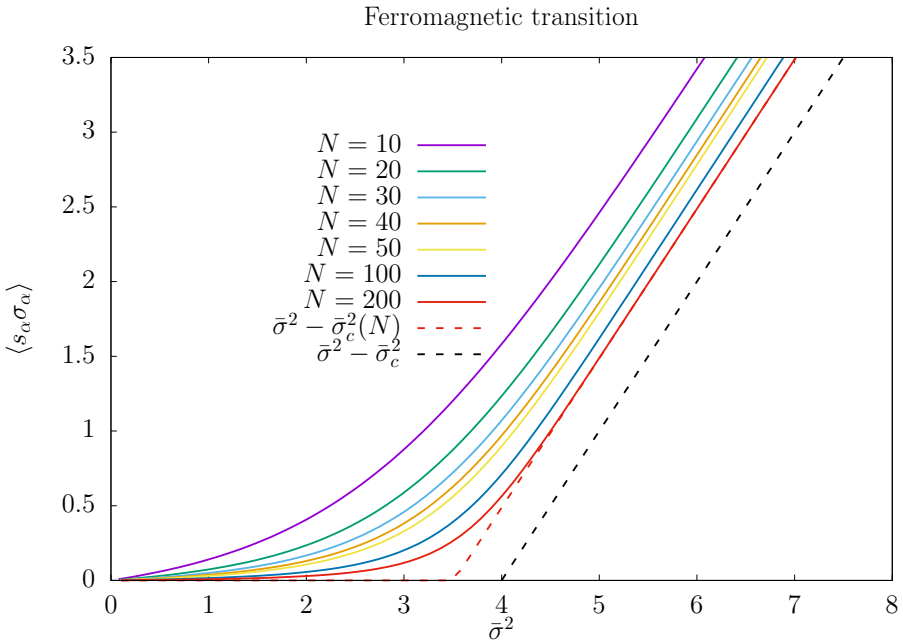

Figure 5. Two-points response function $\langle \hat{s}_\alpha \hat{\sigma}_\alpha \rangle$ with α corresponding to the highest mode, obtained with help of (20), of an RBM having the SD (15) with $\gamma = 0.5$ and $\kappa = 1$. The finite size dependence of the critical $\bar{\sigma}_c^2$ is clearly pronounced. The red dashed curve indicates the finite size behavior when $N = 200$.

this mode, in a change of from a $\mathcal{O}(1/L)$ to a $\mathcal{O}(1)$ behavior. In the queuing terminology, leaving aside the queue corresponding to the reservoir, we expect to see a transition where the queue associated to the smallest service rate absorbs a finite fraction of the total number of clients $\mathcal{O}(L)$ present in the system (not in the reservoir). In fact the two are closely related since we have:

$$\langle \hat{s}_\alpha^2 \rangle_{\text{RBM}} = \frac{2v_\alpha^2}{L} \frac{\partial \log(Z)}{\partial v_\alpha}$$

$$\langle n_\alpha \rangle_{\text{RBM}} = E_\alpha \frac{\partial \log(Z)}{\partial E_\alpha}$$

if v_α is the prior variance of mode $\hat{s}_\alpha$ set to the default value $v_\alpha = 1$ in (1). We then have the relationship (in absence of external fields)

$$\begin{aligned} \langle n_\alpha \rangle_{\text{RBM}} &= L \langle \hat{s}_\alpha \hat{\sigma}_\alpha \rangle_{\text{RBM}} \\ &= \frac{L}{2} \big(\langle \hat{s}_\alpha^2 \rangle_{\text{RBM}} - \frac{1}{L} \big) \geq 0. \end{aligned} \tag{21}$$

The phase transition identified previously actually correspond to an ordinary Bose-Einstein condensation when reinterpreting the n_α as occupation numbers of states α of energy ϵ_α in this last expression, with the identification

$$e^{-\beta \epsilon_\alpha} = E_\alpha$$

and the fugacity ν representing

$$e^{-\beta\nu} = \frac{2z}{\bar{\sigma}^2 L}.$$

Then the critical value $\bar{\sigma}_c$ of $\bar{\sigma}$ previously given in (14) is reintepreted as

$$L\bar{\sigma}_c^2 = \int_{\epsilon_{min}}^{+\infty} d\epsilon \frac{\rho(\epsilon)}{\exp\big[\beta(\epsilon-\nu)\big]-1}\Big|_{\nu=\epsilon_{min}},$$

$L\bar{\sigma}_c^2$ corresponding then to the maximum number of bosons that can be inserted into the system without condensing into the ground state ϵ_{min}.

5 Dynamics of learning

At present we have all the material to setup an "exact" learning method of the Gaussian-spherical RBM, i.e. based on exact response function. The continuous learning equations (7,10,11) given in Section 2.2 are integrated straightforwardly. All the one and two-point correlation functions involved in these equations can be estimated with arbitrary precision in principle, from the previous section. As a result we can generate the deterministic learning trajectories shown on Figure 6. The synthetic data used to train this RBM are generated from a distribution which support is localized in the neighborhood of an ellipsoid of small dimension embedded into a larger dimensional space. As can be checked the modes which emerge eventually align with one of the principal axes of the ellipsoid. The order of arrival is in correspondence with the order of the values of the corresponding principal axes the modes are aligning with. Here the needed time to condense combines the time it takes to align in the right direction and the amplification time of the singular value itself. So as we see in this experiments, 20 out of 100 modes condense which is reflected by the fact that they acquire a finite singular value (close to 2 in this example, top left panel), a finite fraction of the total number of clients (top right panel), or equivalently from (21)a macroscopic variance (bottom right panel). The competition between the modes which condense leads to a non-monotonic behaviour of the fraction of clients or of the variance as new modes show up in the condensate. The mean client number $\langle n_\alpha \rangle_{\mathrm{RBM}}$ associated to mode α is a formal quantity which emerge from the reformulation of the partition function, but cannot be interpreted as an actual degrees of freedom of the learning process.

The critical value of $\bar{\sigma}_c$ of $\bar{\sigma}$ as well as the condensation mechanism itself is defined in principle only in the thermodynamic limit, with a continuous spectrum of the weight matrix. Still, in order to have an estimation of $\bar{\sigma}_c$ in our experiments for a finite size system, we take $z = E_{max}$ and sum over all dominated levels $E < E_{max}$ the integral in (14). As seen in the lower left panel this estimate is meaningful only at the beginning of the learning when the first mode condense, $\bar{\sigma}/\bar{\sigma}_c$ becomes greater than one, but vanish when other modes condense and get close to E_{max}. The evolution of the log likelihood shown shown on the same panel

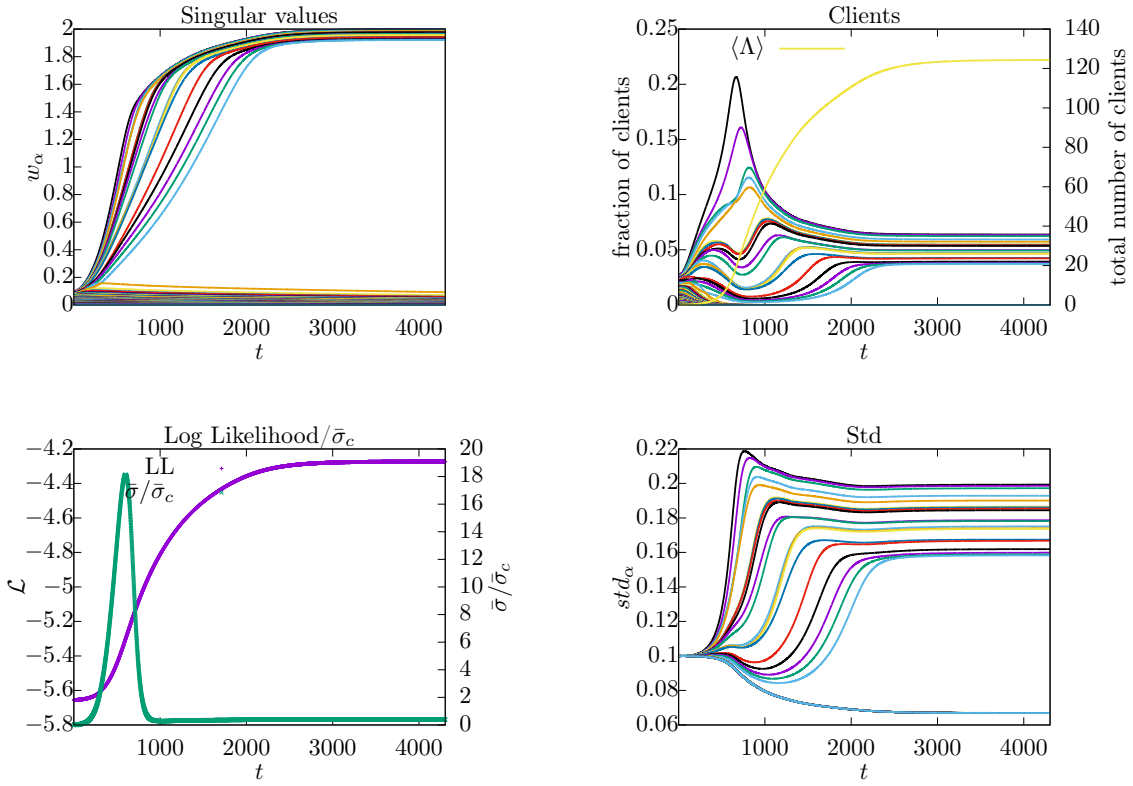

Figure 6. Learning dynamics of an RBM of size $(N_v, N_h) = (100, 100)$ learned on a synthetic dataset of 2000 samples distributed in the neighborhood of a 20-d ellipsoid embedded into a 100-d space. t represents the number of iterations (full batch). The dynamics of the singular values (each color represents one singular value) is shown on the top left panel. Correspondingly the evolution of the number of clients filling the queues is shown on the top right panel or equivalently the variance along each mode on the bottom right. All the modes corresponding to the principal axes of the ellipsoid do eventually condense, while other modes degenerate see their associated singular value, fraction of clients or variance go to zero.

seems to follow very closely the evolution of total mean number of clients shown on the top right panel. We don't have much reason to believe that these two quantities should be analytically related by a simple rescaling, so this coincidence might be presumably incidental.

The scatter plots shown on Figures (7,8) illustrates the ellipsoid shape of the distribution when more than one mode get condensed.

With respect to the thermodynamic analysis given in Section 3.2, the final state of the RBM found by the learning process do not correspond to a continuous bulk of singular values with some given exponent γ. Instead it is a situation where a small number of modes is detached from the bulk of the SD, with mutual distances of order $1/L$, all condensing and capturing in almost equal proportion a finite fraction of the overall variance.

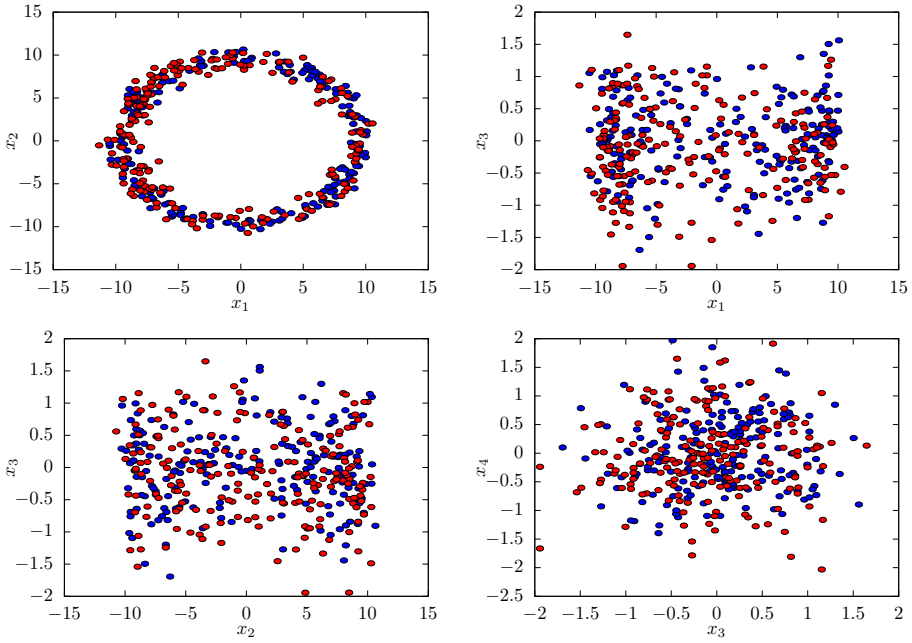

Figure 7. Scatter plot of the training data (blue) and sampled data from the learned RBM (red) projected on the first four svd modes of the weight matrix, for a problem in dimension $N_v = 50$ with two condensed modes.

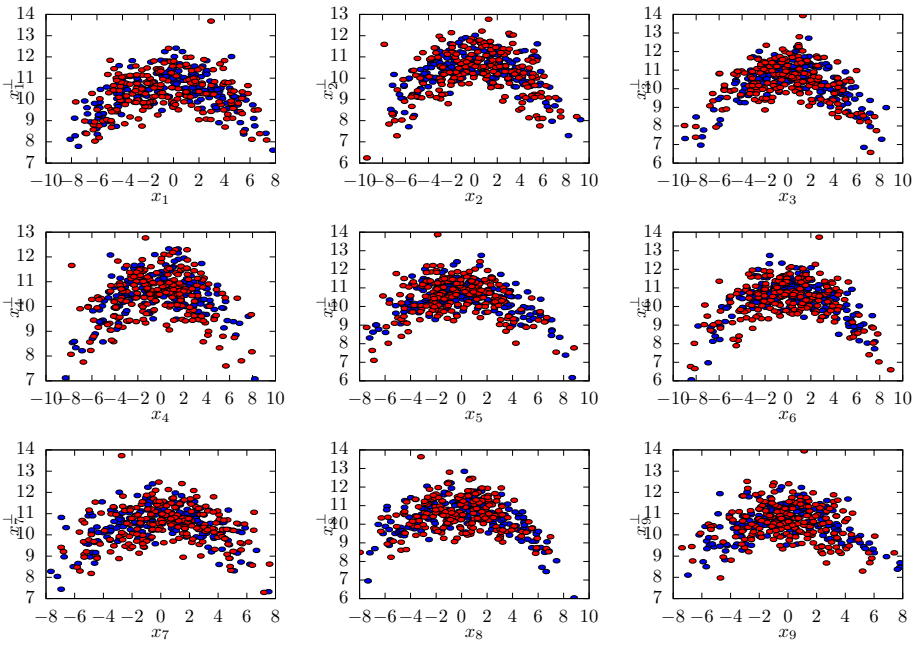

Figure 8. Scatter plot of the training data (blue) and sampled data from the learned RBM (red) projected on the first nine svd modes of the weight matrix (x_i) against the norm of the the orthogonal complement ($x_i^\perp = \sqrt{|x|^2 - x_i^2}$), for a problem in dimension $N_v = 60$ with nine condensed modes.

6 Perspectives

The RBM model presented in Section 4 is clearly useful only from the theoretical point of view, in particular it could serve to test some heuristic learning strategies. When compared to more realistic settings as in [38, 39] we see that this simple model basically exhibits a generic spectral dynamics which account for only one part of the learning process, namely the emergence of modes at the global level leading to the emergence of a simple global manifold supporting the data. Interestingly the Gaussian-spherical RBM is indeed modelling a distribution with a continuous manifold (ellipsoid or portion of ellipsoid if biases are switched on) when more than one mode get condensed. By contrast an RBM with binary latent variables tends to form small spherical clusters to cover the training dataset. To go further we have to concentrate on more realistic and complex situations where the manifold can emerge locally by some piecewise mechanism. This property could maybe be obtained by considering a more general model which can be still solvable using asymptotic response functions of Section 3. First we remark that if we partition the hidden variables into $P = N_h/n$ subsets of size n, and define a spherical constraint on each of these subset, we thereby define a family of models which interpolates between the Gaussian RBM considered so far and an RBM with Ising latent variables when varying n between N_v and 1. Then consider an RBM with N_v and P fixed, but with varying sizes of the partition $\{n_q^h, q = 1, \dots P\}$ with $\sum_q n_q = P$. For the visible variables we consider first an arbitrary basis U partitioned into P subspaces again of sizes $\{n_q^v, q = 1, \dots P\}$, the overall dimension of the visible space being $N_v = \sum_q n_q^v$. We can then define an RBM expressed as a direct product of smaller RBM on this partition directly in the svd mode representation of the global weight matrix:

$$P(\hat{s}, \hat{\sigma}) = \prod_{q=1}^{P} \frac{1}{Z_q} e^{-LE_q(\hat{\boldsymbol{s}}, \hat{\boldsymbol{\sigma}})} \delta\Big(\sum_{\alpha=1}^{n_q^h} \sigma_{q,\alpha}^2 - n_q^h\Big)$$

where now

$$E_q(\hat{\boldsymbol{s}}, \hat{\boldsymbol{\sigma}}) = \sum_{\alpha=1}^{\min(n_q^v, n_q^h)} \big[w_{q,\alpha} \hat{s}_{q,\alpha} \hat{\sigma}_{q,\alpha} + \eta_{q,\alpha} \hat{s}_{q,\alpha} + \theta_{q,\alpha} \hat{\sigma}_{q,\alpha}\big] - \frac{1}{2} \sum_{\alpha=1}^{n_q^v} \frac{\hat{s}_{q,\alpha}^2}{\bar{\sigma}_{q,\alpha}^2},$$

($\bar{\sigma}_{q,\alpha}$ being default variances), and each partition function Z_q can be computed by saddle points approximations. The matrix U as well as the fields $\hat{\eta}$ and $\hat{\theta}$ has to be learned while V is predefined with each mode (q, α) localized on the corresponding subset of hidden variables corresponding to the qth partition.

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[3] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. arXiv 1703.00810, 2017.

[4] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. 2018.

[5] 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.

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

[7] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55(14):1530–1533, 1985.

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

[9] E. Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

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

[11] H. Huang, K.Y. Michael Wong, and Y. Kabashima. Entropy landscape of solutions in the binary perceptron problem. *Journal of Physics A: Mathematical and Theoretical*, 46(37):375002, 2013.

[12] E. Domany and R. Meir. *Layered Neural Networks*, pages 307–334. Springer Berlin Heidelberg, 1991.

[13] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.

[14] F. Krzakala, M. Mézard, F. Sausset, Y.F. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[15] 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.

[16] 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.

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

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

[19] A. Krizhevsky and G. et al. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[20] M. Yasuda and K. Tanaka. Approximate learning algorithm in Boltzmann machines. *Neural Computation*, 21(11):3130–3178, 2009.

[21] K. Cho, A. Ilin, and T. Raiko. Improved learning of Gaussian-Bernoulli restricted Boltzmann machines. In *International conference on artificial neural networks*, pages 10–17. Springer, 2011.

[22] T. Yamashita, M. Tanaka, E. Yoshida, Y. Yamauchi, and H. Fujiyoshii. To be Bernoulli or to be Gaussian, for a restricted Boltzmann machine. In *2014 22nd International Conference on Pattern Recognition*, pages 1520–1525. IEEE, 2014.

[23] V. Nair and G.E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[24] R.D. Hjelm, V.D. Calhoun, R. Salakhutdinov, E.A. Allen, T. Adali, and S.M. Plis. Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. *NeuroImage*, 96:245–260,

2014.

[25] X. Hu, H. Huang, B. Peng, J. Han, N. Liu, J. Lv, L. Guo, C. Guo, and T. Liu. Latent source mining in fmri via restricted Boltzmann machine. *Human brain mapping*, 39(6):2368–2380, 2018.

[26] B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani, and F. Jay. Creating artificial human genomes using generative models. *bioRxiv*, page 769091, 2019.

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

[28] R. Karakida, M. Okada, and S.-I. Amari. Dynamical analysis of contrastive divergence learning: Restricted Boltzmann machines with Gaussian visible units. *Neural Networks*, 79:78 – 87, 2016.

[29] 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.

[30] 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.

[31] J. Tubiana. *Restricted Boltzmann machines: from compositional representations to protein sequence analysis*. PhD thesis, 2018.

[32] A. Barra, G. Genovese, P. Sollich, and D. Tantari. Phase transitions in restricted Boltzmann machines with generic priors. *Physical Review E*, 96(4):042156, 2017.

[33] A. Barra, G. Genovese, P. Sollich, and D. Tantari. Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors. *Physical Review E*, 97(2):022310, 2018.

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

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

[36] A. Maillard, L. Foini, A.L. Castellanos, F. Krzakala, M. Mézard, and L. Zdeborová. High-temperature expansions and message passing algorithms. *arXiv:1906.08479*, 2019.

[37] 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.

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

[39] A. Decelle, G. Fissore, and C. Furtlehner. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *J. Stat. Phys.*, 172:1576–1608, 2018.

[40] T. H. Berlin and M. Kac. The spherical model of a ferromagnet. *Phys. Rev.*, 86:821–835, 1952.

[41] M. R. Evans, S. N. Majumdar, and R. K. P. Zia. Canonical analysis of condensation in factorized steady states. *J.Stat.Phys.*, 123(2):357–390, 2006.

[42] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

[43] L. Pastur. Disordered spherical model. *J.Stat.Phys.*, 27(1):119–151, 1982.

[44] J. M. Kosterlitz, D. J. Thouless, and R. C. Jones. Spherical model of a spin-glass. *Phys. Rev. Lett.*, 36:1217–1220, 1976.

[45] R.J. Baxter. *Exactly Solved Models in Statistical Mechanics*. Academic Press, 1982.

[46] A. Auffinger and W.-K. Chen. Free energy and complexity of spherical bipartite models. *J.Stat. Phys.*, 157(1):40–59, 2014.

[47] J. Baik and J.O. Lee. Free energy of bipartite spherical Sherrington-Kirkpatrick model, 2017.

[48] G. Genovese and D. Tantari. Legendre equivalences of spherical Boltzmann machines. *Journal of Physics A: Mathematical and Theoretical*, 53(9):094001, 2020.

[49] B. Sagan. *The symmetric group*. Springer, 1991.

[50] C. Godrèche and J.M. Luck. Nonequilibrium dynamics of urn models. *Journal of Physics: Condensed*

Matter, 14(7):1601–1615, 2002.

[51] E. Gelenbe and G. Pujolle. *Introduction to Queueing Networks*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[52] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2):248–260, 1975.

[53] J.R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.

A Response functions

All response functions at leading order are expressed as derivatives w.r.t. external fields of ϕ given in (6) taken at the saddle point z_0. We have

$$\begin{aligned}\langle \hat{s}_\alpha \rangle &= \frac{\partial \phi}{\partial \hat{\eta}_\alpha}(z_0) \\ \langle \hat{\sigma}_\beta \rangle &= \frac{\partial \phi}{\partial \theta_\beta}(z_0) \\ \langle \hat{s}_\alpha \hat{\sigma}_\beta \rangle &= \frac{\partial \phi}{\partial \hat{\eta}_\alpha}(z_0)\frac{\partial \phi}{\partial \theta_\beta}(z_0) + \frac{1}{L}\frac{\partial^2 \phi}{\partial \hat{\eta}_\alpha \partial \theta_\beta}(z_0).\end{aligned}$$

We get these as a function of z_0:

$$\begin{aligned}\langle \hat{s}_\alpha \rangle &= \Big(\hat{\eta}_\alpha + \frac{w_\alpha h_\alpha}{z_0 - w_\alpha^2}\Big), \\ \langle \hat{\sigma}_\beta \rangle &= \frac{h_\alpha}{z_0 - w_\alpha^2}, \\ \langle \hat{s}_\alpha \hat{\sigma}_\beta \rangle - \langle \hat{s}_\alpha \rangle \langle \hat{\sigma}_\beta \rangle &= \frac{1}{L} w_\alpha \frac{\partial}{\partial \theta_\beta}\Big(\frac{h_\alpha}{z_0 - w_\alpha^2}\Big) \\ &= \frac{w_\alpha}{L}\Big(\frac{\delta_{\alpha\beta}}{z_0 - w_\alpha^2} - \frac{h_\alpha}{(z_0 - w_\alpha^2)^2}\frac{\partial z_0}{\partial \theta_\beta}\Big) \\ &= \frac{w_\alpha}{L}\left(\frac{\delta_{\alpha\beta}}{z_0 - w_\alpha^2} - \Big(\sum_\gamma \frac{h_\gamma^2}{(z_0 - w_\gamma^2)^3} - \frac{1}{L}\frac{1}{(z_0 - w_\gamma^2)^2}\Big)^{-1}\frac{h_\alpha h_\beta}{(z_0 - w_\alpha^2)^2 (z_0 - w_\beta^2)^2}\right),\end{aligned}$$

where $\partial z_0 / \partial \hat{\theta}_\alpha$ is obtained from the saddle point condition.

B Asymptotic expressions for the condensate

The large deviation function (6) reads (after dropping irrelevant terms) in the continuous formulation

$$\phi(z) = \frac{1}{2}\bar{\sigma}^2 z + \frac{1}{2}\int_0^{E_{\max}} dE \rho(E)\Big(\frac{h(E)^2}{z - E} - \log(z - E)\Big).$$

First we make the change of variable $(z - E_{\max})/E_{\max} \to z$ in the integral representation (5) and change accordingly the definition of the spectral density $\rho(E)dE \to \rho(u = E/E_{\max})du$

and similarly for $\rho(E)h(E)^2$, while $E_{\max}$ is absorbed in the definition of $\bar{\sigma}^2 E_{\max} \to \bar{\sigma}^2$. This leads then to

$$Z(\bar{\sigma}, h) = \frac{E_{\max}}{2i\pi} \int_{-i\infty}^{i\infty} dz \; e^{\frac{L}{2}\phi(z,\bar{\sigma},h)},$$

with

$$\phi(z, \bar{\sigma}, h) = \bar{\sigma}^2(z+1) - \int_0^1 du \rho(u) \Big(\frac{h(u)^2}{z+1-u} - \log(z+1-u) \Big).$$

From the expressions (15,16) and the definition of hypergeometric functions we have

$$\partial_z \phi(z, \bar{\sigma}, h) = \bar{\sigma}^2 - h^2 \Big(\frac{1}{z+1} \Big)^2 F\Big(2, 1+\beta; 2; \frac{1}{z+1}\Big) - \frac{\kappa}{2} \frac{1}{z+1} F\Big(1, 1-\gamma; 2; \frac{1}{z+1}\Big).$$

More general beta distributions with arbitrary exponents would lead as well to hypergeometric functions with different parameters. Our choice leads to more explicit expressions. Indeed, using hypergeometric transformations formulas (See Gradshteyn & Ryzhik) we have:

$$\begin{aligned} F(2, 1+\beta; 2; u) &= (1-u)^{-1-\beta} F(0, 1-\beta; 2; u) \\ &= (1-u)^{-1-\beta} \\ F(1, 1-\gamma; 2; u) &= (1-u)^{\gamma} F(1, 1+\gamma; 2; u) \\ &= \frac{1}{\gamma u} \Big[1 - (1-u)^{\gamma} \Big] \end{aligned}$$

So we finally get for the saddle point equation

$$\begin{aligned} \phi'(z) &= \bar{\sigma}^2 - h^2 \frac{(z+1)^{\beta-1}}{z^{1+\beta}} + \bar{\sigma}_c^2 \Big[\Big(\frac{z}{z+1} \Big)^{\gamma} - 1 \Big], \qquad \text{for} \qquad \gamma \geq 0 \\ &= 0, \end{aligned}$$

with

$$\bar{\sigma}_c^2 = \frac{\kappa}{\gamma}.$$

Upon integration over z we obtain the expression (17) of $\phi(z, \bar{\sigma}, h)$. The expression of the partition function in term of the scaling function $V_{\gamma,\beta}$ is obtained after the rescaling $L^{1/(1+\gamma)} z \to z$.

In absence of external field it becomes obvious that for $\bar{\sigma}_c < \bar{\sigma}$ there is no saddle point solution in the domain $z \geq 0$. This situation has been analyzed in depth in a slightly different context of condensation in zero range processes [41]. In that case, the partition function has a scaling behavior

$$Z_{L,N}(\bar{\sigma}^2) \simeq L^{-\frac{1}{1+\gamma}} V_{\gamma} \Big(L^{\frac{\gamma}{1+\gamma}} (\bar{\sigma}^2 - \bar{\sigma}_c^2) \Big),$$

given in terms of the scaling function (slightly adapting the notation of [41])

$$\begin{aligned} V_\gamma(x) &= \frac{1}{2i\pi}\int_{-i\infty}^{i\infty} du e^{ux+bu^{1+\gamma}}, \\ &= \frac{1}{\pi}\int_0^\infty e^{-b\sin(\gamma\pi/2)u^{\gamma+1}}\cos\Big(b\cos(\gamma\pi/2)u^{\gamma+1}-ux\Big). \end{aligned}$$

Here $L(\bar{\sigma}^2-\bar{\sigma}_c^2)=V_{\text{ex}}$ represents the excess of variance that forces the system to condense on the highest modes. The asymptotic behaviour of V_γ studied in [41] rewrites here:

$$V_\gamma(x) = \begin{cases} \dfrac{b\gamma(\gamma+1)}{\Gamma(1-\gamma)x^{\gamma+2}} & \text{as} \quad x\to\infty \\ \dfrac{b^{-\frac{\gamma}{\gamma+1}}}{(\gamma+1)\Gamma\big(\frac{\gamma}{\gamma+1}\big)} & \text{as} \quad x=0 \\ c_1|x|^{\frac{1-\gamma}{2\gamma}}\exp\big(-c_2|x|^{\frac{\gamma+1}{\gamma}}\big) & \text{as} \quad x\to-\infty \end{cases}$$

with

$$c_1 = \frac{1}{\sqrt{2\pi\gamma}\big(b(\gamma+1)\big)^{\frac{1}{2\gamma}}} \qquad \text{and} \qquad c_2 = \frac{\gamma}{\gamma+1}\big(b(\gamma+1)\big)^{-\frac{1}{\gamma}},$$

leading to

$$Z_{L,N}(\bar{\sigma}^2) \propto \frac{L}{V_{\text{ex}}^{\gamma+2}}$$

in the regime $V_{\text{ex}}=\mathcal{O}(L)$.