Spectral Learning of Restricted Boltzmann Machines

A. Decelle¹ (a), G. Fissore¹,² and C. Furtlehner²

1 LRI, AO team, Bât 660 Université Paris Sud, Orsay Cedex 91405
2 Inria Saclay - Tau team, Bât 660 Université Paris Sud, Orsay Cedex 91405

PACS 02.70.Hm – Spectral methods
PACS 02.30.Zz – Inverse problems
PACS 89.75.-k – Complex systems

Abstract – The restricted Boltzmann machine, an important tool used in machine learning in particular for unsupervised learning tasks, is investigated from the perspective of its spectral properties. Based on empirical observations, we propose a generic statistical ensemble for the weight matrix of the RBM and characterize its mean evolution, with respect to common learning procedures as a function of some statistical properties of the data. In particular we identify the main unstable deformation modes of the weight matrix which emerge at the beginning of the learning and unveil in some way how these further interact in later stages of the learning procedure.

Introduction. – Restricted Boltzmann machines (RBM) [1] constitutes nowadays a common tool on the shelf of machine learning practitioners. It is a generative model, in the sense that it defines a probability distribution, which can be learned to approximate any distribution of data points living in some N-dimensional space, with N potentially large. It also often constitutes a building block of more complex neural network models [2, 3]. The standard learning procedure called contrastive divergence [4] is well documented [5], although being still a not so well understood fine empirical art, with many hyperparameters to tune without much guidelines. At the same time an RBM can be regarded as a statistical physics model, being defined as a Boltzmann distribution with pairwise interactions on a bipartite graph. Similar models have been already the subject of many studies in the 80’s [6–9] which mainly concentrated on the learning capacity, i.e. the number of independent patterns that could be stored in such a model. The second life of neural networks has renewed the interest of statistical physicists for such models. Recent works actually propose to exploit its statistical physics formulation to define mean-field based learning methods using TAP equations [10, 11]. Meanwhile some analysis of its static properties, assuming a given learned weight matrix W, have been proposed [12, 13, 14] in order to understand collective phenomena in the latent representation [15], i.e. the way latent variables organize themselves to represent actual data. One common assumption made in these works is that the weights of W are i.i.d. which as we shall see is unrealistic. Concerning the learning procedure of neural network, many recent statistical physics based analysis have been proposed, most of them within teacher-student setting [16] which imposes a strong assumption on the data, namely that these are generated from a model belonging to the parametric family of interest, hiding as a consequence the role played by the data itself in the procedure. From the analysis of related models [17, 18], it is already a well established fact that a selection of the most important modes of the singular value 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 SVD of the data matrix [19], showing in particular the emergence of each mode by order of importance regarding singular values.

In this work we follow this guideline in the context of a general RBM. We propose to characterize both the learned RBM and the learning process itself by the SVD spectrum of the weight matrix in order to isolate the information content of an RBM. This allows us then to write a deterministic learning equation leaving aside the fluctuations. This equation is subsequently analyzed first in the linear regime to identify the unstable deformation modes of W; secondly at equilibrium assuming the learning is converging, in order to understand the nature of the non-linear

(a) E-mail: aurelien.decelle@lri.fr
interactions between these modes and how these are determined from the input data. In the first section we recall the RBM model and associated learning algorithm. In the second section we show how this algorithm can be described by a generic learning equation. Then we first analyze the linear regime and thereafter we describe what happens with the binary RBM. A set of dynamical parameters is shown to emerge naturally from the SVD decomposition of the weight matrix. The convergence toward equilibrium is analyzed and illustrated later with actual tests on the MNIST dataset.

The RBM and associate 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. The former noted \( s = \{ s_i, i = 1 \ldots N_v \} \) correspond to explicit representation of the data while the latter noted \( \sigma = \{ \sigma_j, j = 1 \ldots N_h \} \) are there to build arbitrary dependencies among the visible units. Overall the RBM defines a joint distribution between visible and hidden units, namely the Boltzmann distribution

\[
p(\mathbf{s}, \mathbf{\sigma}) = \frac{e^{-E(\mathbf{s}, \mathbf{\sigma})}}{Z}
\]

with

\[
E(\mathbf{s}, \mathbf{\sigma}) = - \sum_{i,j} s_i \omega_{ij} \sigma_j - \sum_{i=1}^{N_v} \eta_i s_i - \sum_{j=1}^{N_h} \theta_j \sigma_j
\]

where \( W \) is the weight matrix and \( \eta \) and \( \theta \) are biases, or external fields on the variables. \( Z \) is the partition function of the system. 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 \( W, \eta \) and \( \theta \) such that, new data generated by sampling this distribution should be similar to the input data. The general method to infer the parameters is to maximize the likelihood of the model, where the pdf \( \{1\} \) has first been summed over the hidden variables

\[
\mathcal{L} = \sum_j \log(2 \cosh(\sum_i w_{ij} s_i + \theta_j)) - \log(Z).
\]

Different methods of learning 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 \[10\]. They all correspond to expressing the gradient ascent on the likelihood as

\[
\Delta w_{ij} = \gamma \left( \langle s_i \sigma_j | p(\sigma_j | \mathbf{s}) \rangle_{\text{Data}} - \langle s_i \sigma_j \rangle_{\text{RBM}} \right)
\]

where \( \gamma \) is the learning rate. Similar equations can be derived for the biases. The second term on the rhs is not tractable, and various methods basically differ in their way of estimating this term (Monte-Carlo chains, mean field, TAP...), while the first term is approximated by making use of random mini batches of data at each step.

Deterministic dynamics of the learning. – In order to understand the dynamics of the learning we first project the CD equation \[4\] onto the basis defined by the SVD decomposition of \( W \). In addition we let the learning rate \( \gamma \to 0 \) such that we assume a continuous limit of equation \[4\], and discard stochastic fluctuations usually inherent to the learning procedure. We consider the usual situation where \( N_h < N_v \), which means that the rank of \( W \) is at most \( N_h \). \( W(t) \) represents the learned weight matrix at time \( t \). Let \( \{ w_{\alpha}(t) \in [0, +\infty[ \} \), \( \{ u_{\alpha}(t) \in \mathbb{R}^{N_v} \} \) and \( \{ v_{\alpha}(t) \in \mathbb{R}^{N_h} \} \) denote respectively for \( \alpha = 1, \ldots, N_h \) the singular values, left eigenvectors and right eigenvectors of \( W(t) \), such that the following decomposition \( w_{\alpha}(t) = \sum_{\alpha} u_{\alpha}^* (t) w_{\alpha}(t) e_{\alpha}^* (t) \) holds. Then the continuous version of \(4\) can be recast as follows:

\[
\frac{d w_{\alpha}}{d t} = -\delta_{\alpha, \beta} \frac{d w_{\alpha}}{d t} + \left( 1 - \delta_{\alpha, \beta} \right) \left( w_{\beta}(t) \Omega_{\beta \alpha}(t) + w_{\alpha}(t) \Omega_{\alpha \beta}^h(t) \right) = \langle \sigma \sigma \rangle_{\text{Data}} - \langle \sigma \sigma \rangle_{\text{RBM}}
\]

\[
\Omega_{\alpha \beta}^v(t) = -\Omega_{\beta \alpha}^v \overset{def}{=} \frac{d w_{\alpha}}{d t} A = \frac{2}{w_{\alpha} + w_{\beta}} \left( \frac{d w_{\alpha}}{d t} \right) A_{\alpha \beta} - 2 \frac{w_{\alpha} - w_{\beta}}{w_{\alpha} + w_{\beta}} \frac{d w_{\alpha}}{d t}
\]

\[
\Omega_{\alpha \beta}^h(t) = -\Omega_{\beta \alpha}^h \overset{def}{=} \frac{d w_{\alpha}}{d t} A_{\alpha \beta} = \frac{2}{w_{\alpha} + w_{\beta}} \left( \frac{d w_{\alpha}}{d t} \right) A_{\alpha \beta} + 2 \frac{w_{\alpha} - w_{\beta}}{w_{\alpha} + w_{\beta}} \frac{d w_{\alpha}}{d t}
\]

Here everything is expressed in the reference frame defined by singular vectors of \( W \). \( s_{\alpha} = \sum_{i} u_{\alpha}^* s_i \) and \( \sigma_{\alpha} = \sum_{j} v_{\alpha}^* \sigma_j \) represent spin configurations in this frame. We have introduced the skew-symmetric rotation generators \( \Omega_{\alpha \beta}^v(t) \) of the basis vectors induced by the dynamics. These tell us how the data rotate relatively to this frame. The superscript S,A indicate the symmetric (resp. anti-symmetric) part of the matrix. Note that these equations become singular when some degeneracy occurs in \( W \) because then the SVD decomposition is not uniquely defined. This is not really a problem since we are interested in rotations among non-degenerates modes, the rest corresponding to gauge degrees of freedom. Similar equations can be derived for the fields \( \eta_{\alpha}(t) = \sum_{i} \eta_i(t) u_{\alpha}(t) \) and \( \theta_{\alpha}(t) = \sum_{j} v_{\alpha}(t) \theta_j(t) \) projected onto the SVD modes. At this point we make the
assumption that the learning dynamics is represented by a trajectory of \( \{w_\alpha(t), \eta_\alpha(t), \theta_\alpha(t), \Omega^{\alpha,\beta}_t(t)\} \), while the specific realization of the \( w_\alpha^0 \) and \( v_\alpha^0 \) is considered to be irrelevant, and can be averaged out with respect to some simple distributions, as long as this average is correlated with the data. This means that the decomposition \( \hat{s}_\alpha = \sum_i u_i^\alpha s_i \) of any given sample configuration is assumed also to be kept fixed while averaging. What matters mainly is the strength given by \( w_\alpha \) and the rotation given by \( \Omega^{\alpha,\beta}_t(t) \) of these SVD modes. Assuming for example i.i.d centered normal distribution with respective variance \( 1/N_v \) and \( 1/N_h \) for \( w_\alpha \) and \( v_\alpha \), the empirical term takes the simple form:

\[
\langle s_\alpha \sigma_\beta \rangle_{\text{Data}} = \frac{1}{N_h} \left( s_\alpha (s_\beta w_\beta - \theta_\beta) V \left( \frac{1}{N_h} \sum_\gamma (w_\gamma s_\gamma - \theta_\gamma)^2 \right) \right)_{\text{Data}}
\]

where \( V(x) = \int dy \frac{e^{-y^2/2}}{\sqrt{2\pi}} \text{sech}^2(\sqrt{2}y) \), \( (8) \)

which actually depends on the activation function (an hyperbolic tangent in this case). The main point here is that the empirical term defines an operator which decomposition onto the SVD modes of \( W \) functionally depends solely on \( w_\alpha, \theta_\alpha \) and on the projection of the data on the SVD modes of \( W \). This term is precisely driving the dynamics. The adaptation of the RBM to this driving force is given by the second term which can be well estimated in the thermodynamic limits, as a function of \( w_\alpha, \theta_\alpha \) and \( \eta_\alpha \) alone.

**Linear instabilities.** The first thing to look at with these equations, is the linear regime for which we can actually compute everything. This can be obtained by rescaling all the weights and fields by a common “inverse temperature” \( \beta \) factor and let this go to zero in equations \( (9) \). This limit can be understood by keeping up to quadratic terms in the mean field free energy and should correspond to the first stages of the learning. In this limit, magnetizations \( \langle \mu_\alpha, \mu_\beta \rangle \) of visible and hidden variables have Gaussian fluctuations with covariance matrix

\[
C(\mu_\alpha, \mu_\beta) = \sigma^2_v \begin{bmatrix}
\sigma_v^{-2} & -W^-1 \\
-W^T & \sigma_h^{-2}
\end{bmatrix}^{-1}
\]

with \( \sigma^2_v = 1 \) introduced for sake of generality when considering general linear RBM. To simplify the exposition, we discard the biases of the data and related fields \( \theta_\alpha \) of the RBM. In that case the empirical term in \( (10) \) involves directly the covariance matrix of the data expressed in the frame defined by the SVD modes of \( W \)

\[
\langle s_\alpha \sigma_\beta \rangle_{\text{Data}} = \sigma^2_v \langle s_\alpha s_\beta \rangle_{\text{Data}}.
\]

From \( C(\mu_\alpha, \mu_\beta) \) we get the other terms yielding the following equations:

\[
\frac{dw_\alpha}{dt} = w_\alpha \sigma^2_h \left( \langle s^2 \rangle_{\text{Data}} - \frac{\sigma^2_v}{1 - \sigma^2_v \sigma^2_h w_\alpha^2} \right)
\]

\[
\Omega^{\alpha,\gamma}_t(t) = (1 - \delta_{\alpha,\beta}) \sigma^2_h \left( \frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} \frac{w_\gamma + w_\alpha}{w_\alpha - w_\beta} \right) \langle s_\alpha s_\beta \rangle_{\text{Data}}
\]

Note that these equations are exact for a linear RBM, since they can be derived without any reference to the coordinates of \( u_\alpha \) and \( v_\alpha \) over which we average in the non linear regime. These equations tell us that, during the learning the vectors \( w_\alpha \) (and also \( v_\alpha \)) 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:

\[
w_\alpha^2 = \begin{cases} 
\hat{w}_\alpha^2 - \sigma^2_v & \text{if } \hat{w}_\alpha^2 > \sigma^2_v \\
0 & \text{if } \hat{w}_\alpha^2 \leq \sigma^2_v
\end{cases}
\]

From this we see that the RBM selects the strongest SVD modes in data. The linear instabilities correspond to directions for which the variance of the data is above the threshold \( \sigma^2_v \). This determines the deformations of the weight matrix which can develop during the learning and will eventually interact, following the usual mechanism of non-linear pattern formation like e.g. in reaction-diffusion processes \( (20) \). Other possible deformations are damped to zero. The linear RBM will therefore learn all (up to \( N_h \)) principal components that passed the threshold but it is important to remember that the resulting distribution will still be unimodal. Note that this selection mechanism is already known to occur for linear auto-encoders \( (18) \) or some other similar linear Boltzmann machines \( (17) \). On Fig. 1 we can see the eigenvalues being learned one by one in a linear RBM. For non-linear RBM, or when the system is getting out of the linear regime, we have to develop a proper mean-field theory in order to be able to understand the dynamics and the steady-state regime.

**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 term in \( (10) \) involves higher order statistic of the data as exemplified by \( (8) \) and the Gaussian estimation with \( \sigma^2_v = \sigma^2_h = 1 \) of the RBM response term \( \langle s_\alpha s_\beta \rangle_{\text{RB}} \) is no longer valid. In order to estimate this term in the thermodynamics limits, some assumptions on the form of the weight matrix are at stake. A common assumption consists in to consider i.i.d. random variables for the weights \( w_{ij} \) and this like in \( (13) \) for example generally leads to a Marchenko-Pastur distribution of the singular values of \( W \), which as we shall see in the next section is unrealistic. Instead, based on our experiments this would correspond to
to the noise of the weight matrix, while its information content is better expressed by the presence of SVD modes outside of the bulk. This leads us to write the weight matrix as
\[ w_{ij} = \sum_{\alpha=1}^{K} w_{\alpha} u_{i}^{\alpha} v_{j}^{\alpha} + r_{ij} \tag{9} \]

where the \( w_{\alpha} = O(1) \) are isolated singular values (describing a rank \( K \) matrix), the \( u^{\alpha} \) and \( v^{\alpha} \) are the eigenvectors of the SVD decomposition and the \( r_{ij} = N(0, \sigma^2 / L) \) where \( L = \sqrt{N_h N_v} \) are i.i.d. corresponding to noise. To be consistent with the linear analysis, these modes are assumed to span the (left) subspace corresponding to the part of the empirical SVD above threshold while \( r \) span the complementary space of empirical modes below threshold. We limit the analysis here to the case where \( K \) is finite. This then allows us to assume simple distribution \( p_u \) and \( p_v \) for the components of \( u^{\alpha} \) and \( v^{\alpha} \) considered i.i.d. for instance. This altogether defines our statistical ensemble of RBM to which we restrict ourselves to study the learning procedure. For \( K \) extensive we should instead average over the orthogonal group which would lead to a slightly different mean-field theory \[21,22\]. In the present form our model of RBM is similar to the Hopfield model and recent generalizations \[23\], the pattern 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 \( w_{\alpha} \). Still the analysis in the thermodynamic limit follows the same lines \[7,24\]. In addition to a pair \((q, \bar{q})\) of spin glass parameters over the hidden (resp. visible) nodes, two sets \((m_{\alpha}, \bar{m}_{\alpha}) \approx 1/\sqrt{L} (E(s_{\alpha}), E(\sigma_{\alpha}))\) for \( \alpha = 1, \ldots, K \) of order parameters corresponding to a quenched magnetization toward the SVD directions, are needed to characterize the

retrieval phase. These correspond to average overlap between respectively visible and hidden local magnetizations and the \( K \) singular modes. Assuming a replica-symmetric phase, the saddle-point equations are given by
\[ m_{\alpha} = E_{v,x} \left[ \bar{u}^{\alpha} \tanh \left( \kappa^{-\frac{1}{2}} (\sigma \sqrt{q} v + \sum_{\gamma} (w_{\gamma} \bar{m}_{\gamma} - \theta_{\gamma}) w^{\gamma} ) \right) \right] \]
\[ \bar{m}_{\alpha} = E_{u,x} \left[ u^{\alpha} \tanh \left( \kappa^{-\frac{1}{2}} (\sigma \sqrt{q} u + \sum_{\gamma} (w_{\gamma} m_{\gamma} - \eta_{\gamma}) v^{\gamma} ) \right) \right] \]
\[ q = E_{v,x} \left[ \tanh^2 \left( \kappa^{-\frac{1}{2}} (\sigma \sqrt{q} v + \sum_{\gamma} (w_{\gamma} m_{\gamma} - \eta_{\gamma}) v^{\gamma} ) \right) \right] \]
\[ \bar{q} = E_{u,x} \left[ \tanh^2 \left( \kappa^{-\frac{1}{2}} (\sigma \sqrt{q} u + \sum_{\gamma} (w_{\gamma} m_{\gamma} - \eta_{\gamma}) v^{\gamma} ) \right) \right] \]

where \( \kappa = N_h / N_v \), and where \( E_{u,x} \) and \( E_{v,x} \) denote an average over the Gaussian variable \( x = \mathcal{N}(0,1) \) and the rescaled components \( u \approx \sqrt{N_v} u^{\alpha} \) and \( v \approx \sqrt{N_h} v^{\alpha} \) of the SVD modes. These fixed point equations can be solved numerically to tell us how the variables condensate on the SVD modes within each equilibrium states of the distribution and whether a spin glass phase is present or not. With \( K \) finite and a non-degenerate spectrum the mode with highest singular value dominates the ferromagnetic phase. The phase diagram is in fact similar to the one of the SK model with ferromagnetic coupling, when \( 1/\sigma^2 \) is interpreted as a temperature and \( w_{max} / \sigma^2 \) the ferromagnetic coupling. More interestingly is the presence of meta-stable states which play a role at finite size. Their presence is rather sensitive to the way the average over \( u \) and \( v \) is performed. In this respect the Gaussian distribution with \( u^{\alpha} \) and \( v^{\alpha} \) considered i.i.d. is a special case; all other fixed points associated to lower modes correspond to saddle points. For other distributions with smaller kurtosis, like uniform or Bernoulli, stable fixed points associated to many different single modes or combination of modes can exist. Coming back to the learning dynamics, the first thing which is expected, already from the linear analysis, is that the noise term in (9) vanishes by condensing into a delta function of zero modes. Then the term corresponding to the response of the RBM in (5) is estimated (in absence of bias) in the thermodynamic limits as
\[ \langle s_{\alpha} \sigma_{\beta} \rangle_{\text{RBM}} = \frac{L}{Z_{\text{MF}}} \sum_{q=1}^{C} e^{-F_q m_{\alpha}(q) m_{\beta}(q)} \]

where the index \( q \) run over all MF solutions weighted by their respective free energy
\[ Z_{\text{MF}} \overset{\text{def}}{=} \sum_{q} e^{-F_q} \]

These are the dominant contributions, internal fluctuations given by each fixed point are comparatively of order \( O(1/L) \). 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 \( (f) \) in order to converge. In the case of a multimodal data distribution with many well separated clusters, the SVD modes of \( W \) which will develop are the one pointing in the 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. More precise statements about the phase diagram of the RBM and the behaviour of our dynamical equations will be made in an extended version of this paper, including the dynamics of the fields.

**Tests on the MNIST dataset.** – We illustrate our results on the MNIST dataset. The MNIST dataset is composed of 60000 images of handwritten digits of \( 28 \times 28 \) pixels. It is known that RBMs perform reasonably well on this dataset and therefore we can now interprete in the light of the preceding sections how the learning goes. For the training of the MNIST dataset we use the following parameters. The weights of the matrix \( W \) were initiated randomly from a centered Gaussian distribution with a variance of 0.01 such that the MP bulk do not pass the threshold. The visible fields are initialized to reproduce the empirical mean of the data for each visible variable. The hidden field is put to zero. The learning rate is chosen to be \( \approx 0.01 \). With these parameters we verified that our machine was able to sample digits in a satisfactory way after 20 epochs. Now we can investigate the value of some observable introduced previously. First, we look at the SVD modes of the matrix \( w \) during the learning on Fig. 2. We see that, after seeing only few updates the system has already learned many SVD modes from the data.

On Figure 2a-2c we observe what is expected from the linear regime. Some modes escape from the Marchenko-Pastur bulk of the eigenvalues while other condense down to zero. In particular, we can see that the modes at the beginning of the learning correspond exactly to the SVD modes of the data, see Fig. 3. On this figure, we notice that the modes of the \( W \) matrix are the same as the ones of the data at the beginning of the learning as predicted by the linear theory.

After many epochs, we observe on Fig. 3f that non-linear effects have deformed the SVD modes of \( W \) with respect to the beginning of the learning. We can also look at the evolution of the eigenvalues of \( W \). On Fig. 4 we observe their evolution and when they start to be amplified (or dumped). On the inset, we see how the strongest mode get out of the bulk and increase while the lowest ones are dumped after many epochs. We also observe that the top part of the spectrum of \( W \) appear flattened as compared to empirical SVD spectrum. This presumably favors the expression of many states of similar free energy related to various digit configurations, able to contribute to RBM response term in \( (f) \).

**Discussion.** – The equations obtained for the dynamics and the MF theory that allows us to compute them constitute a phenomenological description of the learning of an RBM. This is assumed to represent a typical learning trajectory in the limit of infinite batch size. 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. The singular values spectrum \( \{\nu_{\alpha}\} \) is playing the main role. The projections \( (\theta_{\alpha}, \hat{\theta}_{\alpha}) \) of the bias onto the eigenmodes of \( W \) are also considered as intrinsic quantities. Finally the rotation vectors \( \{\Omega_{\alpha, \beta}^{h}\} \) give us the rel-
A. Decelle et al.

Fig. 4: Log-log plot of the singular values represented as discrete abscissas (in decreasing order) with their magnitude reported on the ordinates. The RBM contained 400 hidden variables. A cutoff is highlighted by the onset of the linear behaviour and the SVD modes of the data in black. We qualitatively observe that beyond some $\alpha_{\text{thresh}}$ the modes are dumped while before they are amplified. In the inset, the time evolution of the modes 1, 2, 10, 100, 350, 400 during the learning as a function of the number of epochs, we see that for large value of $\alpha$, the modes are decreasing. We observe that the linear cutoff (around $\alpha \approx 50$) seems different from the one observed when going deep into the non-linear regime ($\alpha \approx 250$).

References

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

[2] Hinton G. E. and Salakhutdinov R. R., Science, 313 (2006) 504.

[3] Salakhutdinov R. and Hinton G., Deep Boltzmann machines in proc. of Artificial Intelligence and Statistics 2009 pp. 448–455.

[4] Hinton G. E., Neural computation, 14 (2002) 1771.

[5] Hinton G. E., A Practical Guide to Training Restricted Boltzmann Machines (Springer Berlin Heidelberg, Berlin, Heidelberg) 2012 pp. 599–619.

[6] Hopfield J. J., Proceedings of the National Academy of Sciences of the United States of America, 79 (1982) 2554.

[7] Amit D. J., Gutfreund H. and Sompolinsky H., Annals of Physics, 173 (1987) 30.

[8] Gardner E., EPL (Europhysics Letters), 4 (1987) 481.

[9] Gardner E. and Derrida B., Journal of Physics A: Mathematical and General, 21 (1988) 271.

[10] Gabrièl M., Tramel E. W. and Krzakala F., Training restricted Boltzmann machines via the Thouless-Anderson-Palmer free energy in proc. of Proceedings of the 28th International Conference on Neural Information Processing Systems NIPS’15 2015 pp. 640–648.

[11] Huang H. and Toyoizumi T., Physical Review E, 91 (2015) 050101.

[12] Takahashi C. and Yasuda M., Journal of the Physical Society of Japan, 85 (2016) 034001.

[13] Huang H., Journal of Statistical Mechanics: Theory and Experiment, 2017 (2017) 053302.

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

[15] Monasson R. and Toubiana J., Phys. Rev. Lett., 118 (2017) 138301.

[16] Zdeborová L. and Krzakala F., Advances in Physics, 65 (2016) 453.

[17] Tipping M. E. and Bishop C. M., Neural Comput., 11 (1999) 443.

[18] Bourlard H. and Kamp Y., Biological Cybernetics, 59 (1988) 291.

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

[20] Hohenberg P. C. and Cross M. C., An introduction to pattern formation in nonequilibrium systems (Springer Berlin Heidelberg, Berlin, Heidelberg) 1987 pp. 55–92.

[21] Parisi G. and Potters M., Journal of Physics A: Mathematical and General, 28 (1995) 5267.

[22] Oppen M. and Winther O., Physical Review E, 64 (2001) 056131.

[23] Mézard M., Phys. Rev. E, 95 (2017) 022117.

[24] Amit D. J., Gutfreund H. and Sompolinsky H., Phys. Rev. A, 32 (1985) 1007.