Cascade of phase transitions in the training of Energy-based models

Dimitrios Bachtis[1] **Giulio Biroli**[1] **Aurélien Decelle**[2,3]

Beatriz Seoane[2,3]

[1]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS,
Université PSL, CNRS, Sorbonne Université, Université Paris Cité, F-75005 Paris, France.
[2]Departamento de Física Teórica I, Universidad Complutense, 28040 Madrid, Spain.
[3]Université Paris-Saclay, CNRS, INRIA Tau team, LISN, 91190, Gif-sur-Yvette, France.

Abstract

In this paper, we investigate the feature encoding process in a prototypical energy-based generative model, the Restricted Boltzmann Machine (RBM). We start with an analytical investigation using simplified architectures and data structures, and end with numerical analysis of real trainings on real datasets. Our study tracks the evolution of the model's weight matrix through its singular value decomposition, revealing a series of phase transitions associated to a progressive learning of the principal modes of the empirical probability distribution. The model first learns the center of mass of the modes and then progressively resolve all modes through a cascade of phase transitions. We first describe this process analytically in a controlled setup that allows us to study analytically the training dynamics. We then validate our theoretical results by training the Binary-Binary RBM on real datasets. By using datasets of increasing dimension, we show that learning indeed leads to sharp phase transitions in the high-dimensional limit. Moreover, we propose and test a mean-field finite-size scaling hypothesis. This shows that the first phase transition is in the same universality class of the one we studied analytically, and which is reminiscent of the mean-field paramagnetic-to-ferromagnetic phase transition.

1 Introduction

In recent years, we have witnessed impressive improvements of unsupervised models capable of generating more and more convincing artificial samples [1, 2, 3]. Although energy-based models [4] and variational approaches [5] have been in use for decades, the emergence of generative adversarial networks [6], followed by diffusion models [7], has significantly improved the quality of outputs. Generative models are designed to learn the empirical distribution of datasets in a high-dimensional space, where the dataset is represented as a Dirac-delta pointwise distribution. While different types of difficulties are encounter when training these models, there is a general lack of understanding of how the learning mechanism is driven by the considered dataset. This article explores the dynamics of learning in neural networks, focusing on pattern formation. Understanding how this process shapes the learned probability distribution is complex. Previous studies [8, 9] on the Restricted Boltzmann Machine (RBM) [10] showed that the singular vectors of the weight matrix initially evolve to align with the principal directions of the dataset, with similar results in a 3-layer Deep Boltzmann Machine [11]. Additionally, an analysis using data from the 1D Ising model explained weight formation in an RBM with a single hidden node as a reaction-diffusion process [12]. The main contribution of this work is to demonstrate that the RBM undergoes a series of second-order

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

phase transitions during learning, each corresponding to the acquisition of new data features. This is shown theoretically with a simplified model and on correlated patterns; and confirmed numerically with real datasets, revealing a progressive segmentation of the learned probability distribution into distinct parts and exhibiting second order phase transitions.

2 Related work

The learning behavior of neural networks has been explored in various settings. Early work on deep linear neural networks demonstrated that even simple models exhibit complex behaviors during training, such as exponential growth in model parameters [13, 14]. Using singular value decomposition (SVD) of the weight matrix, researchers revealed a hierarchical learning structure with rapid transitions to lower error solutions. Linear regression dynamics later showed a connection between the SVD of the dataset and the double-descent phenomenon [15]. Similar dynamics were found in Gaussian-Gaussian RBMs [9], where learning mechanisms led to rapid transitions for the modes of the model's weight matrix. In this context, the variance of the overall distribution is adjusted to that of the principal direction of the dataset, while the singular vectors of the weight matrix are aligned to that of the dataset. Unlike linear models, non-linear neural-networks, supervised or unsupervised ones, can not exhibit partition of the input's space. Yet, linear model in general can not provide a multimodal partition of the input space, should it be in supervised or unsupervised context, at difference with non-linear ones.

It was then shown that the most common binary-binary RBMs exhibit very similar patterns at the beginning of learning, transitioning from a paramagnetic to a condensation phase in which the learned distribution splits into a multimodal distribution whose modes are linked to the SVD of the weight matrix [8]. The description of this process motivated the use of RBMs to perform unsupervised hierarchical clustering of data [16, 17]. The succession of phase transitions had been previously observed in the process of training a Gaussian mixture [18, 19, 20], and in the analysis of teacher-student models using statistical mechanics [21, 22]. The latter cases are easier to understand analytically due to the simplicity of the Gaussian mixture. Nevertheless, the learned features are somewhat simpler, as they are mainly represented by the means and variances of the individual clusters. Recently, sequences of phase transitions have been used to explain the mechanism with which diffusion model are hierarchically shaping the mode of the reverse diffusion process [23, 24, 25] and due to a spontaneous broken symmetry [26] after a linear phase converging toward a central fixed-point. The common observation is that the learning of a distribution is, in many cases, obtained by a succession of creation of modes performed through a second order process where the variance in one direction first grow before splitting into two parts, and then the mechanism is repeated. This procedure in particular demonstrate a hierarchical splitting, where the system refined at finer and finer scale of features as it adjust its parameters on a given dataset.

3 Definition of the model

An RBM is a Markov random field with pairwise interactions on a bipartite graph consisting of two layers of variables: visible nodes ($\boldsymbol{v} = \{v_i, i = 1, \ldots, N_\mathrm{v}\}$) representing the data, and hidden nodes ($\boldsymbol{h} = \{h_j, j = 1, \ldots, N_\mathrm{h}\}$) representing latent features that create dependencies between visible units. Typically, both visible and hidden nodes are binary ($\{0, 1\}$), though they can also be Gaussian [27] or other real-valued distributions, such as truncated Gaussian hidden units [28]. For our analytical computations, we use a symmetric representation ($\{\pm 1\}$) for both visible and hidden nodes to avoid handling biases. However, in numerical simulations, we revert to the standard ($\{0, 1\}$) representation. The energy function is defined as follows:

$$E[\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}] = -\textstyle\sum_{ia} v_i W_{ia} h_a - \sum_i b_i v_i - \sum_a c_a h_a, \tag{1}$$

with $\boldsymbol{W}$ the weight matrix and $\boldsymbol{b}$, $\boldsymbol{c}$ the visible and hidden biases, respectively. The Boltzmann distribution is then given by $p[\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}] = Z^{-1}\exp(-E[\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}])$ with $Z = \sum_{\{\boldsymbol{v},\boldsymbol{h}\}} e^{-E[\boldsymbol{v},\boldsymbol{h}]}$ being the partition function of the system. RBMs are usually trained using gradient ascent of the log likelihood (LL) function of the training dataset $\mathcal{D} = \{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(M)}\}$, the LL is then defined as

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) = M^{-1}\sum_{m=1}^{M} \ln p(\boldsymbol{v} = \boldsymbol{v}^{(m)}|\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}) = M^{-1}\sum_{m=1}^{M} \ln \sum_{\{\boldsymbol{h}\}} e^{-E[\boldsymbol{v}^{(m)}, \boldsymbol{h}; \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}]} - \ln Z$$

The computation of the gradient is straightforward and made two terms: the first accounting for the interaction between the RBM's response and the training set, also called *postive term*, and same for the second, but using the samples drawn by the machine itself, also called *negative term*. The expression of the LL gradient w.r.t. all the parameters is given by

$$\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_a \rangle_{\mathcal{D}} - \langle v_i h_a \rangle_{\mathcal{H}}, \quad \frac{\partial \mathcal{L}}{\partial b_i} = \langle v_i \rangle_{\mathcal{D}} - \langle v_i \rangle_{\mathcal{H}} \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial c_a} = \langle h_a \rangle_{\mathcal{D}} - \langle h_a \rangle_{\mathcal{H}}, \tag{2}$$

where $\langle f(\boldsymbol{v},\boldsymbol{h})\rangle_{\mathcal{D}} = M^{-1}\sum_m \sum_{\{\boldsymbol{h}\}} f(\boldsymbol{v}^{(m)},\boldsymbol{h}) p(\boldsymbol{h}|\boldsymbol{v}^{(m)})$ denotes an average over the dataset, and $\langle f(\boldsymbol{v},\boldsymbol{h})\rangle_{\mathcal{H}}$, the average over the Boltzmann distribution $p[\boldsymbol{v},\boldsymbol{h};\boldsymbol{W},\boldsymbol{a},\boldsymbol{c}]$. Most of the challenges in training RBMs stem from the intractable negative term, which has a computational complexity of $\sim \mathcal{O}(2^{\min(N_\mathrm{h},N_\mathrm{v})})$ and lacks efficient approximations. Typically, Monte Carlo Markov Chain (MCMC) methods are used to estimate this term, but their mixing time is uncontrollable during practical learning, leading to potentially out-of-equilibrium training [29].

This work focuses on the initial phase of learning and the emergence of modes in the learned distribution from the gradient dynamics given by Eq. (2). In the following section, we first analytically characterize the early dynamics in a simple setting, showing how it undergoes multiple second-order phase transitions. We then numerically investigate these effects on real datasets.

4 Theory of learning dynamics for simplified high-dimensional models of data

We develop the theoretical analysis by focusing on simplified high-dimensional probability distributions that concentrate around different regions, or *lumps*, in the space of visible variables. Our aim is to analyze how the RBM learns the positions of these lumps, which represent, in a simplified setting, the features present in the data. In order to simplify the analysis, we will consider the Binary-Gaussian RBM (BG-RBM) defined below, yet the same results can be derived for the Binary-Binary RBM (BB-RBM) as shown in the SI B.

4.1 Learning two features through a phase transition

We consider the following simplified setting: we will be using $v_i = \pm 1$ visible nodes, Gaussian hidden nodes and put the biases to zero $\boldsymbol{b} = 0$ and $\boldsymbol{c} = 0$. As a model of data, we consider a Mattis model with a preferred direction $\boldsymbol{\xi}$ for the ground state, following the distribution

$$p_{\mathrm{Mattis}}(\boldsymbol{v}) = \frac{1}{Z_{\mathrm{Mattis}}} \exp\left(\frac{\beta}{2N_\mathrm{v}}\left(\sum_{i=1}^{N_\mathrm{v}} \xi_i v_i\right)^2\right),$$

where $\beta = 1/T$ is the inverse temperature and $\xi_i = \pm 1$ represents a pattern encoded in the model as a Mattis state [30, 31]. In a Mattis model, $\boldsymbol{\xi}$ represents a preferred direction of the model for large values of β, and in this simple case (with only one pattern) there is no need to specify its distribution as long as its elements are ± 1[1]. The Mattis model presents a high-temperature phase with a single mode centred over zero magnetization $m = N_\mathrm{v}^{-1}\sum_i \xi_i \langle v_i\rangle = 0$ for $\beta < \beta_c$ (where $\langle . \rangle$ is the average w.r.t. the Boltzmann distribution) while in the low-temperature regime, $\beta > \beta_c$, the model exhibits a phase transition between two symmetric modes $m = \pm m_0(\beta)$ ($\beta_c = 1$). Henceforth, we shall focus on the regime $\beta > \beta_c$ where the data distribution is concentrated on two lumps. From the analytical point of view, we can compute all interesting quantities in the thermodynamic limit $N_\mathrm{v} \to \infty$. In order to keep the computation simple, we will characterize here the dynamics of the system when performing the learning using a BG-RBM [27] with one single hidden node. In our setting we assume that the distribution of the hidden node is centered in zero (i.e. there is no hidden bias) and that the variance is $\sigma_\mathrm{h}^2 = 1/N_\mathrm{v}$ (we discuss the reason for the scaling in SI A). The distribution is then

$$p_{\mathrm{BG}}(\boldsymbol{h},\boldsymbol{v}) = \frac{1}{Z_{\mathrm{BG}}} \exp\left(\sum_i v_i h w_i - \frac{h^2 N_\mathrm{v}}{2}\right), \; p_{\mathrm{BG}}(\boldsymbol{v}) = \frac{1}{Z} \exp\left[\frac{(\sum_i v_i w_i)^2}{2N_\mathrm{v}}\right].$$

Using this model for the learning, the time evolution of the weights is given by the gradient. With BG-RBM we have that $\langle v_i h\rangle_{\mathcal{H}} = N_\mathrm{v}^{-1}\sum_j w_j \langle v_i v_j\rangle_{\mathcal{H}}$ where the last average is taken over a distribution $p_{\mathrm{BG}}(\boldsymbol{v})$. We can now easily compute the positive and negative term of the gradient w.r.t. the weight

[1] The special case in which all elements $\xi_i = 1$ is the well-known Curie-Weiss model in ferromagnetism.

matrix. For the positive term we obtain that $\langle v_i v_j \rangle_{\mathcal{D}} = \xi_i \xi_j m^2$ where $m = \tanh{(\beta m)}$. The negative term can also be computed in the thermodynamic limit $\langle v_i v_j \rangle_{\mathrm{RBM}} = \tanh(h^* w_i) \tanh(h^* w_j)$ with $h^* = \frac{1}{N} \sum_k w_k \tanh(h^* w_k)$. If we take the limit of a very small learning rate, we can convert the parameter update rule using the gradient into a time differential equation for the parameters of the RBM, where t is the learning time:

$$\frac{dw_i}{dt} = \epsilon \left[\tfrac{1}{N_{\mathrm{v}}} \xi_i \textstyle\sum_k \xi_k w_k m^2 - h^* \tanh(h^* w_i) \right], \tag{3}$$

with ϵ the learning rate [2]. We can analyze two distinct regimes for the dynamics. First, assuming that the weights are small at the beginning of the learning, we get that $h^* = 0$. We can then solve the Eq. (3) in this regime obtaining the evolution of the weights toward the direction $\boldsymbol{\xi}$ by projecting the differential equation on this preferred direction. Defining $U_{\boldsymbol{\xi}} = N^{-1/2} \sum_i \xi_i w_i$, we obtain

$$\tfrac{dU_{\boldsymbol{\xi}}}{dt} = m^2 U_{\boldsymbol{\xi}}, \text{ thus } U_{\boldsymbol{\xi}} = U^0_{\boldsymbol{\xi}} e^{m^2 t}.$$

This illustrates that the weights are growing in the direction of $\boldsymbol{\xi}$ while the projection on any orthogonal direction stays constant. As the weights grow larger, the solution for h^* will depart from zero. Then the correlation between the RBM visible variables starts to grow

$$\langle v_i v_j \rangle_{\mathrm{RBM}} \approx \frac{1}{Z} \textstyle\int dh h^2 w_i w_j \exp\left(-\frac{N_{\mathrm{v}} h^2}{2} + \sum_k \frac{h^2 w_k^2}{2} \right) = w_i w_j \frac{1}{N_{\mathrm{v}} \left(1 - \sum_k w_k^2 / N_{\mathrm{v}} \right)},$$

which means that the susceptibility $\chi = \sum_{i,j} \xi_j \xi_i \langle v_i v_j \rangle_{\mathrm{RBM}}$, that is, the response of the system w.r.t. an external perturbation, diverges when $N_{\mathrm{v}}^{-1} \sum_k w_k^2 \sim 1$, thus exhibiting a *second order phase* transition during the learning. Interestingly, χ diverges as a a power law with a (critical) exponent $\gamma = 1$ (where $N_{\mathrm{v}}^{-1} \sum_k w_k^2$ plays here then the role of the inverse temperature in the standard physical models) thus corresponding to the mean-field universality class [32]. Finally, we can study the regime where the weights are not small. In that case, we can first observe that the evolution of the directions orthogonal to $\boldsymbol{\xi}$ cancel when the weights $\boldsymbol{W}$ align totally with the $\boldsymbol{\xi}$ at the end of the training. Finally, taking $w_i = \xi_i w$, the gradient projected along $\boldsymbol{\xi}$ at stationarity imposes

$$w m^2 = h^* \tanh(h^* w) \text{ and thus } w = \sqrt{\beta} \text{ and } h^* = \sqrt{\beta} m.$$

We confirm the main results of this section numerically in Fig. 1, showing they hold accurately even for moderate values of N_{v}. The sum of the weights grows exponentially, following the magnetization squared (considering the learning rate), and the weights align with the direction $\boldsymbol{\xi}$, while the norm of the weight vector converges towards $\sqrt{\beta}$. Additional analysis details and extended computations for the binary-binary RBM case, which is slightly more involved, are provided in the SI.

4.2 Learning multiple features though a cascade of phase transitions

We consider now the case in which the data are characterized by more than two features. For concreteness, we focus on the case in which the data is drawn from the probability distribution of the Hopfield model [31] with two patterns $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$, using the Hamiltonian $\mathcal{H}_H[\boldsymbol{v}] = -\frac{\beta}{2N_{\mathrm{v}}} \sum_{a=1}^{2} \left(\sum_{i=1}^{N_{\mathrm{v}}} \xi_i^a v_i \right)^2$. The generalization to a larger (but finite) number of patterns is straightforward. Following [33] we consider the case in which the patterns are correlated and defined as: $\boldsymbol{\xi}^1 = \boldsymbol{\eta}^1 + \boldsymbol{\eta}^2$ and $\boldsymbol{\xi}^2 = \boldsymbol{\eta}^1 - \boldsymbol{\eta}^2$; $\boldsymbol{\eta}^1$ is a vector whose first $N_{\mathrm{v}} \frac{1+\kappa}{2}$ components are equal to ±1 with equal probability, and the remaining ones are zero ($0 < \kappa < 1$). Whereas $\boldsymbol{\eta}^2$ is a vector whose last $N_{\mathrm{v}} \frac{1-\kappa}{2}$ components are equal to ±1 with equal probability, and the remaining ones are zero. When $T < 1 - \kappa$ this model is in a symmetry broken phase in which the measure is supported by four different lumps centred in $\pm\boldsymbol{\xi}^1$ and $\pm\boldsymbol{\xi}^2$. Analogously to what was done previously, we now consider a BG-RBM with a number of hidden nodes equal to the number of patterns where again both hidden nodes are centred in zero and have variance $\sigma_h^2 = 1/N_{\mathrm{v}}$. The Hamiltonian is then given by $\mathcal{H}[\boldsymbol{v}, \boldsymbol{h}] = -\sum_{ia} v_i h_a w_{ia} + \sum_a h_a^2 N_{\mathrm{v}}/2$, which corresponds to a Hopfield model [31] with patterns $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$. The analysis presented in the previous section can be generalized to this case (see SI for more details) and one finds the dynamical equations for the evolution of the patterns:

$$\frac{dw_i^a}{dt} = \frac{1}{N_{\mathrm{v}}} \sum_j \langle v_i v_j \rangle_{\mathcal{D}} w_j^a - \frac{1}{N_{\mathrm{v}}} \sum_j \langle v_i v_j \rangle_{\mathrm{RBM}} w_j^a \tag{4}$$

[2] In the rest of the derivation, we will remove it since it can be absorb in a redefinition of the time.

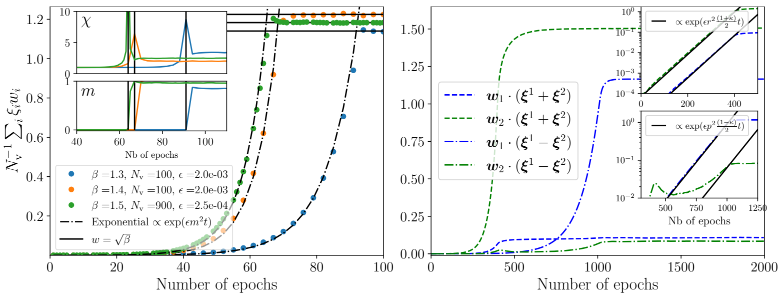

Figure 1: Learning behavior of the BG-RBM with one hidden node, using data from the Mattis model at different inverse temperatures, system sizes and learning rates $\beta, N_{\mathrm{v}}, \epsilon$. The argument of the exponential curves is set to $m^2 \epsilon N_{\mathrm{v}}$, where ϵ is the learning rate. *Inset:* (top) behavior of the susceptibility χ (bottom) magnetization h^* of the learning RBM. The vertical line marks the point at which the susceptibility diverges, indicating the onset of spontaneous magnetization. **Right:** Learning curves for RBMs learning two correlated patterns. The dashed curves represent the weights of the two hidden nodes projected onto $\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$, while the dashed-dotted curves are projected onto $\boldsymbol{\xi}^1 - \boldsymbol{\xi}^2$. *Inset:* Exponential growth during the two phases: top shows growth in the direction $\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$ at a rate $r^2(1+\kappa)/2$, and bottom shows growth in the direction $\boldsymbol{\xi}^1 - \boldsymbol{\xi}^2$ at a rate $p^2(1-\kappa)/2$. The arguments of the exponentials are not adjusted.

As shown in the SI C, $\langle v_i v_j \rangle_{\mathcal{D}} = r^2 \eta_i^1 \eta_j^1 + p^2 \eta_i^2 \eta_j^2$ where r, p are a function of β (and $\beta^{-1} = T < 1 - \kappa$, $r > p$). Note that this factorization of the correlation matrix is precisely its spectral eigendecomposition, which means that $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$ are nothing but the principal directions of the standard principal component analysis (PCA). At the beginning of the training dynamics the RBM is in its high-temperature disordered phase, hence the second term of the RHS of Eq. (4) is zero. The weights $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$ have therefore an exponential growth in the directions $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$, whereas the other components do not evolve. If the initial condition for the weights is very small, as we assume for simplicity, one can then write:

$$\boldsymbol{w}^a(t) = \frac{z^a}{\sqrt{N_{\mathrm{v}}\left(\frac{1+\kappa}{2}\right)}} e^{r^2\left(\frac{1+\kappa}{2}\right)t} \boldsymbol{\eta}^1 + \frac{\tilde{z}^a}{\sqrt{N_{\mathrm{v}}\left(\frac{1-\kappa}{2}\right)}} e^{p^2\left(\frac{1-\kappa}{2}\right)t} \boldsymbol{\eta}^2 \qquad a = 1, 2 \quad ,$$

where we have neglected the small remaining components; z^a and $\tilde{z}^a$ are the projections of the initial condition along the directions $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$. Since $r > p$, on the timescale $(\log N_{\mathrm{v}}) / \left(r^2(1+\kappa)\right)$ the component of the $\boldsymbol{w}^a s$ along $\boldsymbol{\eta}^1$ becomes of order one whereas the one over $\boldsymbol{\eta}^2$ is still negligible. In this regime, the RBM is just like the one we consider in the previous section with a single pattern: the system will align with a single pattern that is given in that case by $\boldsymbol{\eta}^1 \propto \boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$, and it has a phase transition at the time t_I:

$$\frac{e^{2r^2\left[\frac{1+\kappa}{2}\right]t_I}}{N_{\mathrm{v}}} \left((z^1)^2 + (\tilde{z}^1)^2\right) = 1 \ ,$$

At t_I, the RBM learns that the data can be splitted in two groups centred in $\pm\boldsymbol{\eta}^1$, but it does not have yet learned that each one of these two groups consist in two lumps centred in $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ (and respectively $-\boldsymbol{\xi}^1$ and $-\boldsymbol{\xi}^2$). The training dynamics after t_I can also be analyzed: the components of the weight vectors along $\boldsymbol{\eta}^1$ evolve and settle on timescales of order one to a value which is dependent on the initial condition (see the eq. in the SI). In the meanwhile, the components along $\boldsymbol{\eta}^2$ keep growing; at a timescale $(\log N_{\mathrm{v}})/(p^2(1-k))$ (quite larger than t_I in the limit $N_{\mathrm{v}} \to \infty$) they become of order one. In order to analyze easily this regime, let's consider first the simple case in which the initial condition on the weights is such that $\boldsymbol{w}^1(0) \cdot \boldsymbol{\eta}^2 = -\boldsymbol{w}^2(0) \cdot \boldsymbol{\eta}^2$ and $\boldsymbol{w}^1(0) \cdot \boldsymbol{\eta}^1 = \boldsymbol{w}^2(0) \cdot \boldsymbol{\eta}^1$. In this case, one can write $\boldsymbol{w}^1 = A(t)\boldsymbol{\eta}^1 + B(t)\boldsymbol{\eta}^2$ and $\boldsymbol{w}^2 = A(t)\boldsymbol{\eta}^1 - B(t)\boldsymbol{\eta}^2$. The corresponding RBM is a Hopfield model with log likelihood:

$$\sum_a \frac{\left(\sum_i v_i w_i^a\right)^2}{2N_{\mathrm{v}}} = 2A(t)^2 \frac{\left(\sum_i v_i \eta_i^1\right)^2}{2N_{\mathrm{v}}} + 2B(t)^2 \frac{\left(\sum_i v_i \eta_i^2\right)^2}{2N_{\mathrm{v}}}$$

At t_I, when $\left(\frac{1+\kappa}{2}\right) A(t_I)^2 = 1$, one has the first transition in which the RBM measure breaks in two lumps pointing in the direction $\pm\boldsymbol{\eta}^1$, as we explained above. In this regime $B(t)$ is still negligible but keeps increasing with an exponential rate. Using the results of [33], one finds that

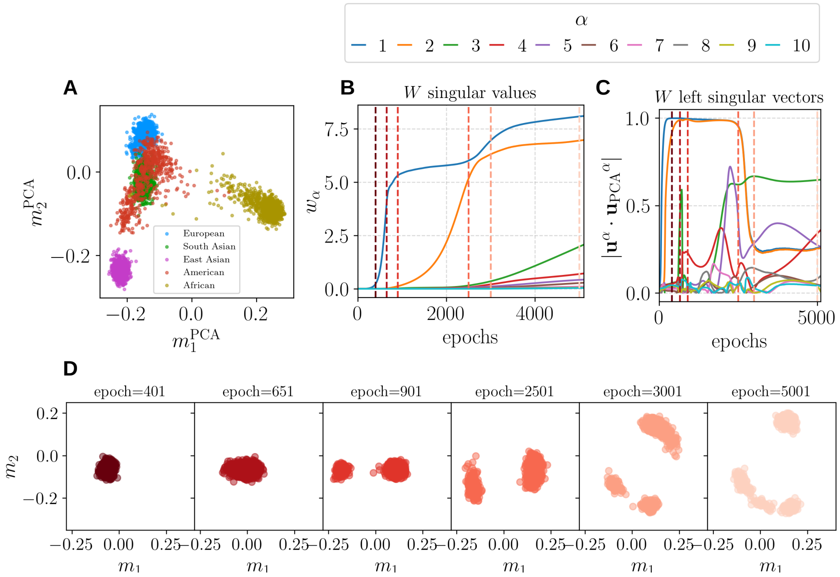

Figure 2: **Human genome dataset.** Progressive coding of the main directions of the dataset when training an RBM with the human genome dataset [34]. In A, we show the dataset projected along the first two principal components of the dataset $\boldsymbol{\eta}_\alpha$ with $\alpha = 1, 2$, and $m_\alpha^{\text{PCA}} = \boldsymbol{\eta}_\alpha \cdot \boldsymbol{x}^{(d)}/\sqrt{N_{\text{v}}}$, with $\boldsymbol{x}^{(d)}$ referring to the different entries in the dataset, i.e. an human individual. Points are colored according to the individual continental origin. In B, we show the evolution of the singular values w_α of the RBM weight matrix $\boldsymbol{W}$ as a function of the number of training epochs, and in C, we show the scalar product of the corresponding singular vectors $\boldsymbol{u}_\alpha$ with the corresponding PCA component $\boldsymbol{\eta}_\alpha$. In D, we show the magnetization of the samples generated by the model at different epochs, projected along the first two eigenvectors of $\boldsymbol{W}$, which shows that the specialization of the model occurs through the progressive encoding of the main modes of the data in $\boldsymbol{W}$.

when $\frac{1-\kappa}{2}B(t_{II})^2 = 1$, a second phase transition takes place. This defines a time t_{II} at which the probability measure of the RBM breaks from two lumps to four lumps, each one centred around one of the four directions $\pm\boldsymbol{\xi}^1, \pm\boldsymbol{\xi}^2$. We have considered a special initial condition, but the phenomenon we found is general. In fact, for any initial condition one can show that the dynamical equations have an instability on the timescale t_{II}, which generically induces the second symmetry breaking transition. On Fig. 1, right panel, we illustrate the exponential growth as described by the theory, toward the two directions. In the SI 4.2, we show how these phase transitions are in very good agreement with previous work [9, 8] and how the phase space is split during training time. At the end of the training, the patterns are given by $\boldsymbol{w}^1 = \xi^1$ and $\boldsymbol{w}^2 = \xi^2$ modulo a rotation in the subspace spanned by $\boldsymbol{\xi}^{1,2}$, since the likelihood is invariant by rotations in this subspace. In fact, we often found that the patterns are not perfectly aligned because we are not forcing the weights to be binary. This analysis can naturally extend to more than two patterns, typically resulting in a cascade of phase transitions. In this process, the RBM progressively learns the data features, starting from a coarse-grained version (just the center of mass) and gradually refining until all patterns are learned. The analysis done on the BG-RBM can of course be repeated on the BB-RBM (as is done for the case with one mode in the SI B). The main difference at that level between the two models is that in order to have a retrieval phase, the BB-RBM needs to encode the patterns on an extensive number of hidden nodes (proportional to N_{v}), while the BG-RBM needs only as many patterns as hidden nodes. Both models can match perfectly the dataset in the limit $N_{\text{v}} \to \infty$, but we might encounter discrepancies for finite size. However, when dealing with real datasets, by construction, the BG-RBM can not reproduce higher-order correlations and therefore is less interesting than the BB case.

5 Numerical Analysis

In the previous sections, we examined the learning process in simplified setups, in order to be able to develop an analytical treatment. In particular, we have shown analytically in a simple setting how the weight matrix is shaped by the patterns present in the dataset and how the learning process dynamics is triggered by the PCA components (the $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$ in the previous example) and not by the learning of the encoded patterns ($\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$). Moreover, we have shown that each time the RBMs learn a new direction, the susceptibility of the system diverges with a precise power law everytime the RBM learns a new direction, which is also associated with the development of new modes in the probability measure. In this section we will also show that the insights gained from this simplified analysis are also applicable to understanding the learning process of a Binary-Binary RBM (BB-RBM) with many hidden nodes trained with real data sets. The details about the training procedure are given in SI E.1. For this purpose, we will consider 3 real data sets: (i) The Human Genome Dataset (HGD), (ii) MNIST and (iii) CelebA, see details in the SI D. To show the occurrence of bonafide phase transitions, it is important to show the effect of increasing the system size (which transforms cross-overs in sharp transitions in the large $N_{\rm v}$ limit). We will therefore resize these data sets in different dimensions by adjusting their resolution, i.e. by changing $N_{\rm v}$ while maintaining comparable statistical properties. Detailed information about the scaling process can be found in the SI D.

In real training processes, the machine is expected to gradually learn different patterns $\boldsymbol{\xi}^\alpha$, from the data, as described in the previous sections. However, the identification of these patterns and their relationship to the statistical properties of the dataset remains unclear. Previous research [9, 8, 35] has shown that RBM training begins with the stepwise encoding of the most significant principal components of the dataset, $\{\boldsymbol{\eta}^\alpha\}$, which are the eigenvectors of the sample covariance matrix with the highest eigenvalues, on the SVD decomposition of its weight matrix $W_{ia} = \sum_\alpha w_\alpha u_i^\alpha \bar{u}_a^\alpha$, where $\boldsymbol{u}^\alpha \in \mathbb{R}^{N_{\rm v}}$ and $\bar{\boldsymbol{u}}^\alpha \in \mathbb{R}^{N_{\rm h}}$ denote the left and right singular vectors corresponding to the singular value w_α. These vectors form orthonormal bases in $\mathbb{R}^{N_{\rm v}}$ and $\mathbb{R}^{N_{\rm h}}$ respectively, where the index α ranges from 1 to $\min(N_{\rm v}, N_{\rm h})$ and the singular values w_α are arranged in descending order. At the beginning of the learning process, the left singular vectors, $\boldsymbol{u}^\alpha$, gradually align α-to-α to $\boldsymbol{\eta}^\alpha$. This is consistent with the analytical results in our simple setting in the previous sections. In analogy to the mean-field magnetic models proposed in the previous section, the role of decreasing temperature is played by the increasing magnitude of the singular value w_α^2, associated with each mode α, and should lead to a series of phase transitions where the RBM measure splits into increasingly more and more modes. We show in Fig. 2 that these phenomena are at play by focusing on the evolution of the SVD of the RBM weight matrix when trained with the HGD dataset.

In panel A we show the first two principal components of the dataset, which highlights its strong multimodal structure, as several distant clusters appear (in this case, they are related to the continental origin of the individuals at hand). In Fig. 2–B we show the sharp and sudden increases of the singular values w_α, as expected from our theoretical analysis, and in Fig. 2–C the evolution of the scalar product between $\boldsymbol{u}_\alpha$ and $\boldsymbol{\eta}_\alpha$ as a function of the number of training epochs. Different colors indicate different values of α. As expected, the modes are progressively expressed during training, and the first two singular vectors match the two principal directions of the dataset for a while. This last figure also shows us that the alignment with the PCA is only temporary (a limitation of current theoretical approaches), as the machine finds better patterns to encode the data as training progresses.

The progressive splitting of the RBM measure during the training dynamics is shown in Fig. 2–D, for which we use $N_{\rm s} = 1000$ independent samples generated with the model trained up to a different number of epochs (the colors refer to the same epochs highlighted with vertical lines in Figs. 2–B and C). For visualization, we show the samples projected onto the right singular vectors of $\boldsymbol{W}$, the *magnetizations* $m_\alpha = \boldsymbol{v} \cdot \boldsymbol{u}^\alpha / \sqrt{N_{\rm v}}$ with $\alpha = 1, 2$. At the beginning of training, the data points are essentially Gaussian distributed, and the growth of w_1 over 4 is related to the splitting of the data into two different clusters on the m_1 axis, and the emergence of w_2 is related to a second splitting on the m_2 axis. At this stage of training, the projections along all subsequent directions are Gaussian distributed as they are the result of a sum of random numbers (fixed by the random initialization of the weight matrix). This progressive splitting is crucial to express the diversity of the dataset shown in Fig. 2–A, and can be successfully used to extract relational trees to cluster data points, as recently shown in Ref. [16]. The details about the numerical analysis are given in E.2.

At the beginning of training, when only a singular value has been expressed, and thus $\boldsymbol{W} \approx w\boldsymbol{u}\bar{\boldsymbol{u}}^\top$, the transition of the feature encoding process is analogous to the phase transition from the paramagnetic to the ferromagnetic phase in the Mattis model mentioned above with pattern $\boldsymbol{u}$. The detailed justification can be found in SI F. Our analysis allows us to define an effective temperature, linked to the eigenmode of $\boldsymbol{W}$ as $\beta = w^2/16$. Now, since the critical temperature of the Mattis model is $\beta_c = 1$, we can show that the BB-RBM will condensate when the first eigenmode of the model

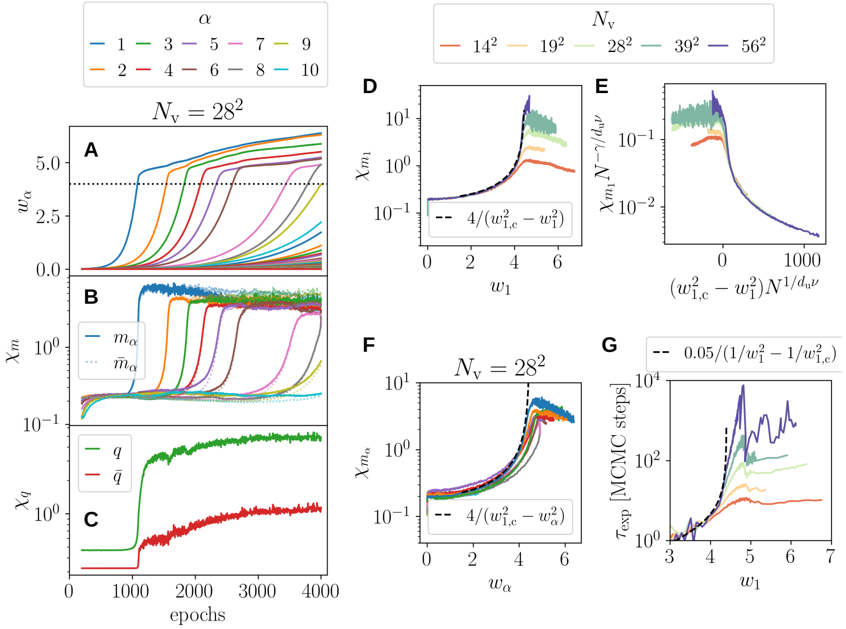

Figure 3: **Traning with the MNIST dataset.** In A we show the evolution of the singular values of the RBM's coupling matrix $\boldsymbol{W}$ as a function of the training time. In B we show the evolution of the susceptibilities associated with the magnetizations along the right singular vectors of $\boldsymbol{W}$, $m_\alpha = \langle \boldsymbol{u}_\alpha \cdot \boldsymbol{v} \rangle / N_{\mathrm{v}}$. In both figures, we consider the standard $N_{\mathrm{v}} = 28^2$ MNIST dataset, different colors refer to different modes. In C we show the susceptibility associated with the overlaps q and $\bar{q}$ between visible and hidden variables. In D we show the susceptibility of the first mode as a function of the first singular value w_1 obtained with trainings on MNIST data scaled to different system sizes above and below $L = 28$. The numerical curves are compared with the theoretical expectation using the Mattis model in Eq. (6) using $w_{1,\mathrm{c}} = 4.45$. The same data are shown in E, scaled using the mean-field finite-size scaling ansatz of Eq. (7). In F, we show the first 10 modes' susceptibilities χ_{m_α} as a function of their corresponding singular value w_α and compare them with the theoretical curve in D. In G, we show the MCMC relaxation time of the machines trained with different N_{v} datasets as a function of w_1, together with the theoretical expectation for local moves in dashed lines.

reaches $w_c = 4$, see SI F. In a real training, we also have visible $\boldsymbol{b}$ and hidden bias $\boldsymbol{c}$ which could easily change the model towards a random field Mattis model, which leads us to expect a slightly higher critical point but a very similar ferromagnetic phase transition, and in particular, it should not change the transition's mean-field universality class.

To show that there is a cascade of transitions and that what was found for the HGD also holds for other datasets, we now train the RBM with the MNIST dataset. In Fig. 3–A we plot the evolution of the singular values w_α along the training, which clearly show the progressive encoding of patterns. The progressive splitting of the RBM measure into clusters and the presence of a phase transition can be monitored by measuring the variance of the distribution of the visible magnetizations m_α along the α-th mode or the analogous hidden magnetizations $\bar{m}_\alpha = \boldsymbol{h} \cdot \bar{\boldsymbol{u}}^\alpha / \sqrt{N_{\mathrm{h}}}$ obtained using the hidden units. The variance of the magnetization multiplied by the number of variables used to compute it and β, is related to the *magnetic susceptibility* via the fluctuation dissipation theorem, which means that

$$\chi_m = N_{\mathrm{v}} \left(\langle m^2 \rangle - \langle m \rangle^2 \right) = T \, \mathrm{d} \langle m \rangle / \mathrm{d}h, \tag{5}$$

here $\langle \cdot \rangle$ refers to the equilibrium measure with respect to RBM's Gibbs measure $p(\boldsymbol{v}, \boldsymbol{h})$, in practice estimated as the average over N_{s} independent MCMC runs. It is well known that the magnetic susceptibility should diverge in the vicinity of a second order phase transition and that such growth in only limited by the overall system size $N = \sqrt{N_{\mathrm{v}} N_{\mathrm{h}}}$ in finite systems. These phenomena indeed takes place also in the RBM. We show in Fig. 3–B the evolution of the χ_ms obtained using the magnetizations obtained along the different modes α of $\boldsymbol{W}$. As anticipated, the susceptibility χ_{m_1}

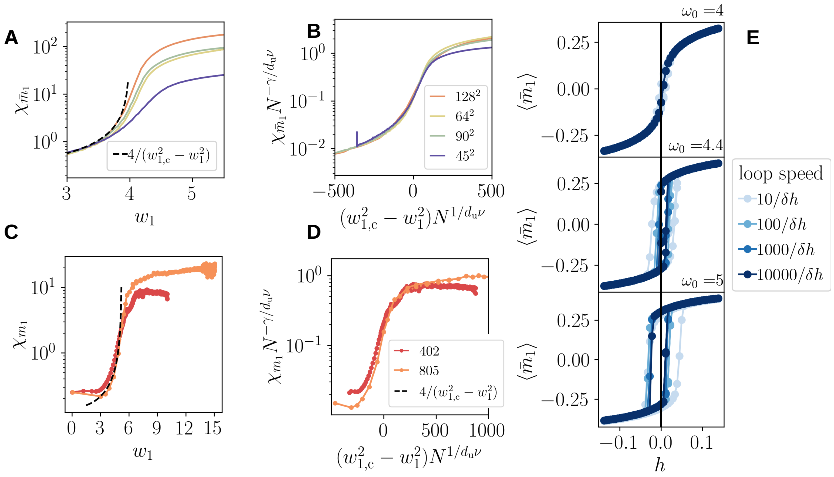

Figure 4: **Training with the CELEBA and HGD datasets:** In A, we plot the hidden susceptibility for different system sizes in the CELEBA dataset, with dashed lines indicating the expected divergence at $w_{1,c} = 4$. In B, we show the mean-field FFS associated with the first transition using mean-field exponents. In C and D, we present the visible susceptibility for the first phase transition in the HGD dataset, using $w_{1,c} = 5.25$ for scaling. In E, typical hysteresis in the low-temperature phase is illustrated for CELEBA (128×128), similar to the mean-field Ising model in external fields.

associated to the magnetization m_1 along the first mode, sharply grows as w_1 approaches 4, but it is more remarkable that this behavior is not only restricted to the first mode, but it is also reproduced by the subsequent modes in a step-wise process. According to the mapping between the low-rank RBM and the Mattis (or equivalently, the Curie-Weiss) model, we should expect that our χ_m, at least for the first mode $\alpha = 1$ should behave as

$$\chi_m \sim \frac{4}{\beta_\mathrm{c} - \beta} = \frac{4}{w_\mathrm{c}^2 - w^2}, \tag{6}$$

when approaching the critical point, which is equivalent to stating that the critical exponent is $\gamma = 1$. Here, the factor 4 in the numerator is related to the fact that the susceptibility obtained with $\{0, 1\}$ variables is 4 times the standard one obtained with Ising spins, and that $\beta = w^2/16$. In Fig. 3–D, we show the susceptibility associated with the first mode as a function of w_1 using RBMs trained with MNIST data rescaled to different dimensions. As mentioned earlier, the growth of the susceptibility is limited by the system size N_v. However, if we look at increasingly larger sizes, we can observe the growth over several decades. This shows that at the transition we observe the Mattis/Curie-Weiss behavior of Eq. 6, as shown in the black dashed line, where the only adjustable parameter was the critical point $w_{1,\mathrm{c}} = 4.45$ (i.e. there is no adjustable pre-factor).

One of the crucial tests to ensure that a finite-size transition is a bona fide phase transition is to study its behavior by changing the number of degrees of freedom. One of the standard tools to do this is to make use of the so-called finite-size scaling (FSS) ansatz, motivated by renormalization group arguments [36, 37, 38]. Mean-field models follow a modified FSS ansatz which was first studied in [39]. In particular, the FSS ansatz for the susceptibility is

$$\chi_m^N(\beta) = N^{\frac{\gamma}{\nu d_\mathrm{u}}} \phi\left(N^{\frac{1}{\nu d_u}} |\beta - \beta_\mathrm{c}|\right), \tag{7}$$

with $\phi(\cdot)$ a size-independent scaling function, $N = \sqrt{N_\mathrm{v} N_\mathrm{h}}$ is the effective size of our model and $\gamma = 1$, $\nu = 1/2$ and $d_u = 4$ as expected in the mean-field universality class. We test this ansatz in Fig. 3–E showing that it does succeed to scale the finite-size data in the critical region, especially in the largest system sizes, which confirms both the mean-field universality class and the prevalence of the transition in the thermodynamic limit. In Fig. 4-A and B, and Fig. 4-C and D, we show that the indicators of a phase transitions–growth of the susceptibility and its mean-field finite size scaling–also holds for CelebA and HGD datasets. Finally, a final piece of evidence of the existence of a phase transition is presented in Fig. 4-E, where we show that after the continuous transition has taken place, one can induce a discontinuous transition and hysteresis effects by applying a field in the direction of the learned pattern, in full agreement with what observed for standard phase transitions, the details of the analysis can be found in SI E.2.2, and further theoretical insights into the relationship between hysteresis and discontinuous transitions can be found in G.

All the discussion so far has been mainly concerned with the first phase transition, when the RBM learns the first mode. But we have discussed in Fig. 3–B that an entire sequence of step-wise phase transitions occurred in the rest of the $\boldsymbol{W}$-matrix modes. In Fig. 3–F we show each of these mode susceptibilities χ_{m_α} as a function of their corresponding singular value. They show extremely similar divergent behavior with respect to the mode $\alpha = 1$, with an apparent slight variation of the critical point for each mode, although all appear to remain close to the predicted $w_\alpha \sim 4$, suggesting that the subsequent transitions may be of similar mean-field nature. Exceeding second-order phase transitions has a very strong impact on the overall quality of the training, in particular on the quality of the log-likelihood gradient estimated by MCMC dynamics. Indeed, second-order transitions are associated with a well-known arresting effect, known as *critical slowing-down* behavior, by which the thermalization times diverge with the correlation length $\xi^z \propto |\beta - \beta_{\mathrm{c}}|^{-\nu z}$, where z is the dynamical critical exponent, which is 2 for local and non-conserved order parameter moves in mean-field, making the thermalization of large systems extremely difficult in practice. We show that our exponential relaxation times diverge exactly as predicted in Fig. 3–G. This has a significant impact on the quality of models trained with maximum likelihood approaches, as these methods rely on MCMC to estimate gradients. It is therefore expected that MCMC mixing times increase sharply each time a mode is coded, which can be prohibitive for clustered and high-dimensional datasets. Recent studies have shown that pre-training a low-rank RBM using other methods (and thus bypassing the initial phase transitions) can be very effective in improving the models in clustered datasets [40]. However, we emphasize that the cascade of phase transitions described in this paper occurs regardless of the training scheme or whether the Markov chains reach equilibrium. This is discussed further in SI H.

Extensions and limitations– All these results can be studied in detail for RBMs thanks to the fact that we can analytically deal with the Hamiltonian. However, our results can be extended to the case of Deep Boltzmann Machines, where previous works have also computed the phase diagram, which are also based on the SVD decomposition of the weighing matrices [11], but also in diffusion models where phase transitions linked to the learning has also been described [24, 26]. It therefore stands to reason that similar phenomena occur with even more complex models such as Convolutional EBM, but where it is not clear how the parameters of the model can be decomposed. A first test would be to see what the projection of the generated data would look like in the different phases of learning.

6 Conclusions

In this paper, we first characterized the learning mechanism of RBMs using a simplified setting with a dataset provided by a simple teacher model. We used two examples: one with two symmetric clusters and another with four correlated clusters. Our results show that the learning dynamics identify modes by exponential growth in the directions of the clusters dominated by the variances of these clusters. The theory predicts the timing of the first phase transitions and agrees well with [8]. Numerically, we have confirmed the existence of a cascade of phase transitions associated with the growing modes w_α and accompanied by divergent susceptibility. Finite-size scaling suggests that these transitions are critical and fall into the class of mean-field universality. This set of phase transitions likely goes beyond RBMs and offers insights into learning mechanisms, particularly for generative models. These transitions have significant implications for both training and understanding the learned features. During training, each transition is associated with a divergent MCMC relaxation time, which requires careful handling to properly train the model. In addition, the hysteresis phenomenon ensures that the learning trajectory involves second-order phase transitions, which are beneficial for tracking the emergence of modes in the learned distribution. However, changing parameters (such as the local bias) could lead to first-order transitions that are detrimental to sampling and could explain the ineffectiveness of parallel tempering in the presence of temperature changes. In practice, our analysis shows that the principal directions of the weight matrix contain valuable information for understanding the learned model.

7 acknowledgments

The authors would like to thank J. Moreno-Gordo for his work and discussions during the initial phase of this project. The authors (A.D. and B.S.) acknowledge financial support by the Comunidad de Madrid and the Complutense University of Madrid (UCM) through the Atracción de Talento program (Refs. 2019-T1/TIC-13298), and to project PID2021-125506NA-I00 financed by the Ministerio de Economía y Competitividad, Agencia Estatal de Investigación (MICIU/AEI /10.13039/501100011033), and the Fondo Europeo de Desarrollo Regional (FEDER, UE).

References

[1] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681–4690, 2017.

[2] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. Advances in Neural Information Processing Systems, 32, 2019.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021.

[4] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

[7] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pages 2256–2265. PMLR, 2015.

[8] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Thermodynamics of restricted boltzmann machines and related learning dynamics. Journal of Statistical Physics, 172:1576–1608, 2018.

[9] Aurélien Decelle, Giancarlo Fissore, and Cyril Furtlehner. Spectral dynamics of learning in restricted boltzmann machines. Europhysics Letters, 119(6):60001, 2017.

[10] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In Artificial intelligence and statistics, pages 448–455. PMLR, 2009.

[11] Yuma Ichikawa and Koji Hukushima. Statistical-mechanical study of deep boltzmann machine given weight parameters after training by singular value decomposition. Journal of the Physical Society of Japan, 91(11):114001, 2022.

[12] Moshir Harsh, Jérôme Tubiana, Simona Cocco, and Remi Monasson. 'place-cell'emergence and learning of invariant data with restricted boltzmann machines: breaking and dynamical restoration of continuous symmetries in the weight space. Journal of Physics A: Mathematical and Theoretical, 53(17):174002, 2020.

[13] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120, 2013.

[14] Andrew M Saxe, James L McClellans, and Surya Ganguli. Learning hierarchical categories in deep neural networks. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 35, 2013.

[15] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. Neural Networks, 132:428–446, 2020.

[16] Aurélien Decelle, Beatriz Seoane, and Lorenzo Rosset. Unsupervised hierarchical clustering using the learning dynamics of restricted boltzmann machines. Physical Review E, 108(1):014110, 2023.

[17] Jorge Fernandez-De-Cossio-Diaz, Thomas Tulinski, Simona Cocco, and Rémi Monasson. Replica symmetry breaking and clustering phase transitions in undersampled restricted boltzmann machines. 2024.

[18] Kenneth Rose, Eitan Gurewitz, and Geoffrey C Fox. Statistical mechanics and phase transitions in clustering. Physical review letters, 65(8):945, 1990.

[19] David Miller and Kenneth Rose. Hierarchical, unsupervised learning with growing via phase transitions. Neural Computation, 8(2):425–450, 1996.

[20] Tony Bonnaire, Aurélien Decelle, and Nabila Aghanim. Cascade of phase transitions for multiscale clustering. Physical Review E, 103(1):012105, 2021.

[21] N Barkai and Haim Sompolinsky. Statistical mechanics of the maximum-likelihood density estimation. Physical Review E, 50(3):1766, 1994.

[22] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 601–608. IEEE, 2016.

[23] Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. 2023(9):093402, oct 2023.

[24] Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. arXiv preprint arXiv:2402.18491, 2024.

[25] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. arXiv preprint arXiv:2402.16991, 2024.

[26] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. Advances in Neural Information Processing Systems, 36, 2024.

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[28] 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.

[29] Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. Equilibrium and non-equilibrium regimes in the learning of restricted boltzmann machines. Advances in Neural Information Processing Systems, 34:5345–5359, 2021.

[30] DC Mattis. Solvable spin systems with random interactions. Physics Letters A, 56(5):421–422, 1976.

[31] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8):2554–2558, 1982.

[32] Giorgio Parisi and Ramamurti Shankar. Statistical field theory. 1988.

[33] Francisco A Tamarit and Evaldo MF Curado. Pair-correlated patterns in hopfield model of neural networks. Journal of statistical physics, 62:473–480, 1991.

[34] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature, 526(7571):68, 2015.

[35] Aurélien Decelle and Cyril Furtlehner. Restricted boltzmann machine: Recent advances and mean-field theory. Chinese Physics B, 30(4):040202, 2021.

[36] Kurt Binder. Finite size scaling analysis of ising model block distribution functions. Zeitschrift für Physik B Condensed Matter, 43:119–140, 1981.

[37] Daniel J Amit and Victor Martin-Mayor. Field theory, the renormalization group, and critical phenomena: graphs to computers. World Scientific Publishing Company, 2005.

[38] John Cardy. Finite-size scaling. Elsevier, 2012.

[39] E Brézin. An investigation of finite size scaling. Journal de Physique, 43(1):15–22, 1982.

[40] Nicolas Béreux, Aurélien Decelle, Cyril Furtlehner, Lorenzo Rosset, and Beatriz Seoane. Fast, accurate training and sampling of restricted boltzmann machines. arXiv preprint arXiv:2405.15376, 2024.

[41] Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. Physical Review E, 95(2):022117, 2017.

[42] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[43] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. CoRR, abs/1710.10196, 2017.

[44] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In Proceedings of the 25th international conference on Machine learning, pages 1064–1071, 2008.

[45] Frank den Hollander. Metastability under stochastic dynamics. Stochastic Processes and their Applications, 114(1):1–26, 2004.

[46] Anton Bovier. Metastability, pages 177–221. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[47] Paul M Chaikin, Tom C Lubensky, and Thomas A Witten. Principles of condensed matter physics, volume 10. Cambridge university press Cambridge, 1995.

[48] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8):1771–1800, 2002.

[49] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. Advances in Neural Information Processing Systems, 32:5232–5242, 2019.

A Binary-Gauss RBM

We add some technical details to the derivation of the dynamical process. We first recall the definition of general BG-RBM first. BG-RBMs consist of a bipartite model where the visible nodes $\boldsymbol{v}$ are binary (in our case ± 1) and the hidden nodes are Gaussians. The Hamiltonian follows the same expression as the BB-RBM, Eq. 1. The hidden nodes are Gaussian random variables centred in zero with variance σ_h^2. In our analysis where we use only one or few hidden nodes, it is important that the variance scales as the inverse of the system's size $\sigma_h^2 = 1/N_{\mathrm{v}}$ in order for the Hamiltonian of the system to be an extensive property (proportional to N_{v}). This scaling is crucial for the analysis: if the energy term scales as the system's size, then it is possible to have two distinct phases: for small weight magnitude, the system is disordered (it does not polarize in any particular direction) while for large values of $\boldsymbol{w}$ the system will polarize toward one of the direction encoded in the weight matrix. A more detailed description can be found in [30, 41].

Now, recall that we consider the Mattis model, biased toward a pattern $\boldsymbol{\xi}$ for generating the dataset

$$p_{\mathrm{Mattis}}(\boldsymbol{v}) = \frac{1}{Z_{\mathrm{Mattis}}} \exp\left(\frac{\beta}{2N_{\mathrm{v}}}\left(\sum_{i=1}^{N_{\mathrm{v}}} \xi_i v_i\right)^2\right)$$

where β is the inverse temperature $\beta = 1/T$ and $\xi_i = \pm 1$ represents a potential pattern direction. The Mattis model presents a high-temperature phase with a single model centred over zero magnetization $m = N_{N_{\mathrm{v}}}^{-1} \sum_i v_i = 0$ for $\beta < \beta_c$ while in the low-temperature regime, $\beta > \beta_c$, the model exhibits a phase transition between two symmetric modes $m = \pm m_0(\beta)$. From the analytical point of view, we can compute all interesting quantities in the thermodynamics limit $N_{\mathrm{v}} \to \infty$. The RBM's distribution is given by

$$\begin{aligned} p_{RBM}(\boldsymbol{h}, \boldsymbol{v}) &= \frac{1}{Z_{RBM}} \exp\left(\sum_i v_i h w_i - \frac{h^2 N_{\mathrm{v}}}{2}\right) \\ p_{RBM}(\boldsymbol{v}) &= \frac{1}{Z} \exp\left(\frac{(\sum_i v_i w_i)^2}{2N_{\mathrm{v}}}\right) \end{aligned}$$

Using this model for the learning, the time evolution of the weights is given by the gradient. With BG-RBM we have that

$$\begin{aligned} \langle v_i h \rangle &= \sum_{\{\boldsymbol{v}\}} \int dh v_i h p(\boldsymbol{v}, h) = \sum_{\{\boldsymbol{v}\}} \int dh v_i h p(h|\boldsymbol{v}) p(\boldsymbol{v}) && (8) \\ &= \frac{1}{N_{\mathrm{v}}} \sum_{\{\boldsymbol{v}\}} v_i \sum_j v_j p(\boldsymbol{v}) w_j = \frac{1}{N_{\mathrm{v}}} \sum_j w_j \langle v_i v_j \rangle_p && (9) \end{aligned}$$

where the last average is taken over a distribution $p(\boldsymbol{v})$. We can now easily compute the positive and negative term of the gradient w.r.t. the weight matrix. For the positive term, assuming that $\beta > 1$, we obtain that

$$\begin{aligned} \langle v_i v_j \rangle_{\mathcal{D}} &= \frac{1}{Z_{\mathrm{Mattis}}} \int dm \sum_{\{\boldsymbol{v}\}} v_i v_j \exp\left(-\beta N_{\mathrm{v}} \frac{m^2}{2} + m\beta \sum_k \xi_k v_k\right) \\ &= \frac{1}{Z_{\mathrm{Mattis}}} \int dm \tanh(\beta \xi_i m) \tanh(\beta \xi_j m) \exp\left(-\beta N_{\mathrm{v}} \frac{m^2}{2} + \sum_k \log 2\cosh(\beta \xi_k m)\right) \end{aligned}$$

Evaluating the saddle point of the argument of the exponential (which is the same as the one for the partition function) we have that

$$\langle v_i v_j \rangle_{\mathcal{D}} = \xi_i \xi_j m^2 \text{ where } m = \tanh(\beta m)$$

The negative term can also be computed in the thermodynamic limit

$$\begin{aligned}
\langle v_i v_j \rangle_{\text{RBM}} &= \frac{1}{Z_{\text{RBM}}} \int dh \sum_{\boldsymbol{v}} v_i v_j \exp\left(\sum_k v_k h w_k - \frac{h^2 N_{\text{v}}}{2} \right) \\
&= \int dh \frac{1}{Z_{\text{RBM}}} \tanh(h w_i) \tanh(h w_j) \exp\left(\sum_k \log\left[2\cosh(h w_k)\right] - \frac{h^2 N_{\text{v}}}{2} \right) \\
&= \tanh(h^* w_i) \tanh(h^* w_j) \text{ with } h^* = \frac{1}{N_{\text{v}}} \sum_k w_k \tanh(h^* w_k)
\end{aligned}$$

where the last line is obtain by taking the saddle point of the integral over h, (h^* corresponding to the extremum). We can now express the gradient as

$$\begin{aligned}
\frac{dw_i}{dt} &= \frac{1}{N_{\text{v}}} \xi_i \sum_k \xi_k w_k m^2 - \frac{1}{N_{\text{v}}} \sum_k w_k \tanh(h^* w_k) \tanh(h^* w_i) \\
&= \frac{1}{N_{\text{v}}} \xi_i \sum_k \xi_k w_k m^2 - h^* \tanh(h^* w_i)
\end{aligned}$$

Assuming first that the weights are small we get that $h^* = 0$. We can solve the gradient's equations in this regime. In such case, the only solution for the saddle point equation of the RBM is given by $h^* = 0$ and we can see that the solution of the evolution of the weight is global toward the direction $\boldsymbol{\xi}$ by projecting the differential equation on the preferred direction. Defining $U_{\boldsymbol{\xi}} = N_{\text{v}}^{-1/2} \sum_i \xi_i w_i$, we obtain

$$\frac{dU_{\boldsymbol{\xi}}}{dt} = m^2 U_{\boldsymbol{\xi}} \text{ thus } U_{\boldsymbol{\xi}} = U_{\boldsymbol{\xi}}^0 e^{m^2 t}.$$

This shows that the weights are growing in the direction of $\boldsymbol{\xi}$ while the projection on any orthogonal direction $\boldsymbol{\phi}^\alpha$ stays constant:

$$\boldsymbol{\phi}^\alpha \cdot \frac{d\boldsymbol{w}}{dt} = \frac{m^2}{N_{\text{v}}} (\boldsymbol{\phi}^\alpha \cdot \boldsymbol{\xi})(\boldsymbol{w} \cdot \boldsymbol{\xi}) = 0 \text{ since } \boldsymbol{\phi}^\alpha \cdot \boldsymbol{\xi} = 0$$

When the weights grow larger, the solution for h^* will depart from zero. The correlation of the learning RBM then starts to grow

$$\langle v_i v_j \rangle_{\text{RBM}} \approx \frac{1}{Z} \int dh h^2 w_i w_j \exp\left(-\frac{N_{\text{v}} h^2}{2} + \sum_k \frac{h^2 w_k^2}{2} \right) = w_i w_j \frac{1}{N_{\text{v}} \left(1 - \sum_k w_k^2 / N_{\text{v}}\right)}$$

$$\chi = \sum_{i,j} \xi_j \xi_i \langle v_i v_j \rangle_{\text{RBM}} \approx \left(\sum_i \xi_i w_i \right)^2 \frac{1}{N_{\text{v}} \left(1 - \sum_i w_i^2 / N_{\text{v}}\right)}$$

and diverges when $N_{\text{v}}^{-1} \sum_i w_i^2 \sim 1$, therefore exhibiting a second order phase transition during the learning. Finally, we can study the regime where the weights are not small. In that case, we can first observe that the evolution of the directions orthogonal to $\boldsymbol{\xi}$, $\boldsymbol{\phi}^\alpha$ are given by

$$\sum_i \phi_i^\alpha \frac{dw_i}{dt} = \frac{m^2}{N_{\text{v}}} \sum_i \phi_i^\alpha \xi_i \sum_k \xi_k w_k - \sum_i \phi_i^\alpha h^* \tanh(h^* w_i) = -\sum_i \phi_i^\alpha h^* \tanh(h^* w_i)$$

which will cancel if the weight $\boldsymbol{W}$ aligns totally with the $\boldsymbol{\xi}$. Finally, taking $w_i = \xi_i w$, the gradient projected along $\boldsymbol{\xi}$ at stationarity imposes

$$w m^2 = h^* \tanh(h^* w) \text{ and thus } w = \sqrt{\beta} \text{ and } h^* = \sqrt{\beta m}$$

B Binary-Binary RBM

The RBM sharing both discrete binary variables on the visible and hidden nodes is by far the most commonly used. In particular, using binary nodes in the hidden layer instead of the Gaussian distribution allows the model to potentially fit any order correlations of the dataset. In this section, we review how the learning dynamics translate to this case, using for simplicity binary $\{\pm 1\}$ variables. In order to obtain an interesting behavior in this phase of the learning, it is important to consider a particular parametrization of the RBM. We consider that all hidden nodes share the same weight. This is important to be able to have a recall phase transition in the model. We therefore have the following Hamiltonian

$$\mathcal{H} = -\frac{1}{N_{\rm h}} \sum_i v_i w_i \sum_a h_a \tag{10}$$

where $N_{\rm h} = \alpha N_{\rm v}$ is the number of hidden nodes h_a of the system and the vector $\boldsymbol{W}$ correspond to the weight shared across all the hidden nodes. In this model, we can now compute the positive and negative of the gradient. The first one is given by

$$\begin{aligned}
\frac{1}{N_h}\langle v_i \sum_a h_a \rangle_{\mathcal{D}} &= \frac{1}{Z_{\rm Mattis}} \sum_{\boldsymbol{v}} \int dm v_i \exp\left(-\frac{\beta m^2 N_{\rm v}}{2} + m\beta \sum_j \xi_j v_j\right) \frac{1}{N_{\rm h}} \sum_a \tanh\left[N_{\rm h}^{-1} \sum_j w_j v_j\right] \\
&= \frac{1}{Z_{\rm Mattis}} \sum_{\boldsymbol{v}} \int dm d\tau v_i \exp\left(-\frac{\beta m^2 N_{\rm v}}{2} + m\beta \sum_j \xi_j v_j\right) \delta(\tau - N_{\rm h}^{-1} \sum_j w_j v_j) \tanh(\tau) \\
&= \frac{1}{Z_{\rm Mattis}} \sum_{\boldsymbol{v}} \int d\tau d\bar{\tau} dm v_i \exp\left(-\frac{\beta m^2 N_{\rm v}}{2} + m\beta \sum_j \xi_j v_j\right) \tanh(\tau) e^{i\tau\bar{\tau} - iN_{\rm h}^{-1}\bar{\tau}\sum_j w_j v_j} \\
&= \frac{1}{Z_{\rm Mattis}} \int dm d\tau d\bar{\tau} e^{-\beta m^2 N_{\rm v}/2 + i\tau\bar{\tau}} \tanh(\xi_i m\beta - iN_{\rm h}^{-1}\bar{\tau} w_i) \tanh(\tau) \\
&\qquad \times \exp\left(\sum_j \log\cosh\left[\xi_j \beta m - iN_{\rm h}^{-1}\bar{\tau} w_j\right]\right)
\end{aligned}$$

finding the saddle point of the argument in the exponential, we obtain

$$\frac{1}{N_h}\langle v_i \sum_a h_a \rangle_{\mathcal{D}} = \xi_i \tanh(\beta m) \tanh\left(\frac{m}{N_{\rm h}} \sum_j \xi_j w_j\right) = \xi_i m \tanh\left(\frac{m}{N_{\rm h}} \sum_j \xi_j w_j\right)$$

The same type of computation can be done for the negative term, we found that

$$\frac{1}{N_h}\langle v_i \sum_a h_a \rangle_{RBM} = \xi_i \tau \tanh(w_i \tau)$$

$$\tau = \tanh\left(\frac{1}{N_h} \sum_j w_j \tanh(w_j \tau)\right)$$

Again, in the small coupling regime (or at the beginning of the learning), when $N_{\rm h}^{-1} \sum_j w_j^2 \ll 1$, we have that $\tau = 0$. In such case, the gradient over the weight matrix is given by

$$\frac{dw_i}{dt} = \xi_i m \tanh\left(\frac{m}{N_{\rm h}} \sum_j \xi_j w_j\right)$$

following the same approach as in the main text, we project the weights on the unit vector $\boldsymbol{u}_1 = \boldsymbol{\xi}/\sqrt{N_{\rm v}}$, $U_{\boldsymbol{\xi}} = \boldsymbol{u}_1 \boldsymbol{W}$, which gives

$$\frac{dU_{\boldsymbol{\xi}}}{dt} = \sqrt{N_{\rm v}} m \tanh\left(\frac{m\sqrt{N_{\rm v}}}{N_{\rm h}} U_{\boldsymbol{\xi}}\right)$$

We can integrate this equation, obtaining the solution

$$\sinh\left(\frac{m}{\sqrt{N_{\rm v}}\alpha} U_{\boldsymbol{\xi}}(t)\right) = \sinh\left(\frac{m}{\sqrt{N_{\rm v}}\alpha} U_{\boldsymbol{\xi}}(0)\right) \exp\left(\frac{m^2 t}{\alpha}\right)$$

$$U_{\boldsymbol{\xi}}(t) = U_{\boldsymbol{\xi}}(0) \exp\left(\frac{m^2 t}{\alpha}\right)$$

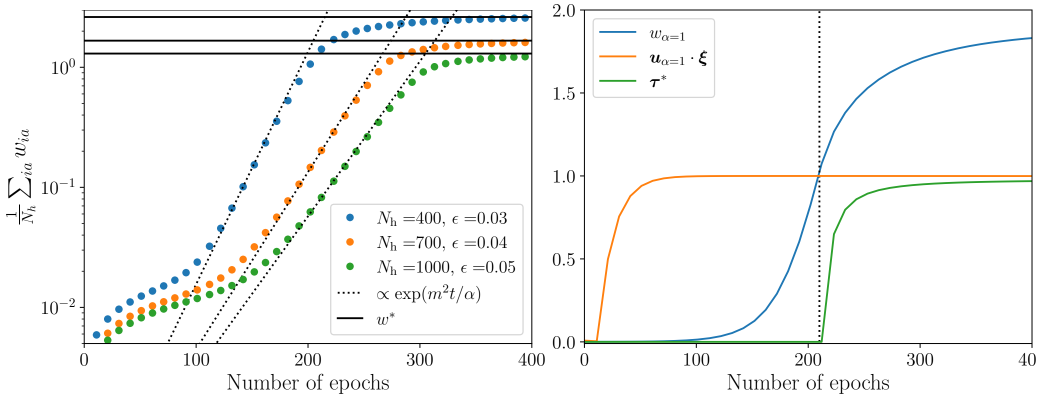

Figure 5: **Left:** learning behavior of the Binary-Binary RBM, using data from the Mattis model. The different curves correspond to systems of size $N_\mathrm{v} = 900$ at inverse temperature $\beta = 1.4$ with learning rate $\epsilon = 0.03, 0.04, 0.05$ and $N_\mathrm{h} = 400, 700, 1000$ respectively. The argument of the exponential curves are not adjusted but set to $m^2\epsilon/\alpha$. **Right:** we illustrate the RBM's dynamics in the binary-binary case with $\beta = 1.4$ and $N_\mathrm{v} = 900$, $N_\mathrm{h} = 400$. First the eigenvector $\boldsymbol{u}^{\alpha=1}$ aligns itself with the pattern $\boldsymbol{\xi}$. Then, the eigenvalue $w_{\alpha=1}$ grows exponentially until reaching saturation and when it crosses the value 1, the system develops a spontaneous magnetization.

where the second line is obtained in the very large N_v limit. Again we have an exponential growth in the first steps of the learning. At the end of the learning, the weights again align in the direction of $\boldsymbol{\xi}$. This can be checked by the fact that the positive term of always orthogonal to any vector orthogonal to $\boldsymbol{\xi}$, and thus the simplest option for the gradient projected in those direction is to be orthogonal to $\boldsymbol{\xi}$. Taking $\boldsymbol{W} = \boldsymbol{u}_1 w$, we obtain

$$m \tanh\left(mw/\alpha\right) = \tau \tanh(w\tau)$$
$$\tau = \tanh(w \tanh(w\tau)/\alpha)$$

The solution can be found numerically by solving the fixed point equation on τ, and measuring the magnetization of the dataset. In Fig.5 we illustrate our results in the same dataset as in the section 4.1, taking the Mattis model with $N_\mathrm{v} = 900$, $\beta = 1.4$, varying the learning rate and the number of hidden nodes.

C Learning with correlated patterns

In this part we detail how the learning goes when considering a pair of correlated patterns. As described in 4.2, the pairs of patterns are defined as

$$\boldsymbol{\xi}^1 = \boldsymbol{\eta}^1 + \boldsymbol{\eta}^2 \text{ and } \boldsymbol{\xi}^2 = \boldsymbol{\eta}^1 - \boldsymbol{\eta}^2$$

where $\boldsymbol{\eta}^1$ is a vector whose first $N_\mathrm{v}\frac{1+\kappa}{2}$ components are equal to ± 1 with equal probability and the remaining ones are zero. The other vector $\boldsymbol{\eta}^2$ has its last $N_\mathrm{v}\frac{1-\kappa}{2}$ components equal to ± 1 with equal probability and the rest are 0; we also have that $\kappa \in [0, 1]$. When $\kappa = 1$, both patterns $\boldsymbol{\xi}^{1,2}$ are equal, while otherwise different but correlated. In particular $\mathbb{E}_{\boldsymbol{\eta}^1,\boldsymbol{\eta}^2}[\boldsymbol{\xi}^1\boldsymbol{\xi}^2] = N_\mathrm{v}\kappa$. Following the results of [33], it is possible to compute the saddle point equations for the magnetization. The general form is given by

$$m_1 = \frac{1}{N_\mathrm{v}}\sum_i \xi_i^1 \tanh\left(\beta m_1 \xi_i^1 + \beta m_2 \xi_i^2\right)$$
$$m_2 = \frac{1}{N_\mathrm{v}}\sum_i \xi_i^2 \tanh\left(\beta m_1 \xi_i^1 + \beta m_2 \xi_i^2\right)$$

This system has been solved in [33] and exhibits the following properties. When $T > 1 + \kappa$, the system is in the paramagnetic regime and $m_1 = m_2 = 0$. When the temperature is lowered and lies in $1-\kappa < T < 1+\kappa$, the solution is given by the pair retrieval $m_1 = m_2 = m = \frac{1+\kappa}{2}\tanh(2\beta m)$. Finally, when $T < 1-\kappa$, the system condensates on the following solution

$$
\begin{aligned}
m_1 &= \frac{1+\kappa}{2}\tanh\left(\beta(m_1+m_2)\right) + \frac{1-\kappa}{2}\tanh\left(\beta(m_1-m_2)\right) \\
m_2 &= \frac{1+\kappa}{2}\tanh\left(\beta(m_1+m_2)\right) - \frac{1-\kappa}{2}\tanh\left(\beta(m_1-m_2)\right)
\end{aligned}
$$

where basically the system either condensates toward one of the pattern $\boldsymbol{\xi}^{1,2}$, while the other magnetization has some non-zero value due to the correlation.

We can use the thermodynamics properties of this model to study how the learning of the RBM should behave in the regime $T < 1-\kappa$. In order to use this model as generating the dataset, we need to compute the correlations $\langle s_i s_j \rangle$. The model presents four fixed points, all equally probable:

$$
\begin{aligned}
&(m_1, m_2) = (m^+, m^-) \text{ and its symmetric case } (m_1, m_2) = (-m^+, -m^-) \\
&(m_1, m_2) = (m^-, m^+) \text{ and its symmetric case } (m_1, m_2) = (-m^-, -m^+)
\end{aligned}
$$

where $m^+ > m^- > 0$. Therefore, writing $r = \tanh(\beta(m^+ + m^-))$ and $p = \tanh(\beta(m^+ - m^-))$ we have that

$$
\begin{aligned}
\langle v_i v_j \rangle_{data} &= \frac{1}{4}\left(\sum_{(m_1,m_2)} \left[(\eta_i^1 + \eta_i^2)\tanh(\beta(m_1+m_2))\right]\left[(\eta_j^1 + \eta_j^2)\tanh(\beta(m_1+m_2))\right] \right) \\
&= \eta_i^1 \eta_j^1 r^2 + \eta_i^2 \eta_j^2 p^2
\end{aligned}
$$

because the cross terms $\eta_i^1 \eta_j^2$ are canceled when changing $(m_1 = m^+, m_2 = m^-)$ to $(m_1 = m^-, m_2 = m^+)$. At this point, it is possible to write the gradient at the linear order and project it toward both direction $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$. Denoting $U^a_{\boldsymbol{\eta}^1} = \boldsymbol{\eta}^1 \cdot \boldsymbol{w}^a$ and $U^a_{\boldsymbol{\eta}^2} = \boldsymbol{\eta}^2 \cdot \boldsymbol{w}^a$, we get

$$
\begin{aligned}
\frac{dU^a_{\boldsymbol{\eta}^1}}{dt} &= r^2 \frac{1+\kappa}{2} U^a_{\boldsymbol{\eta}^1} \\
\frac{dU^a_{\boldsymbol{\eta}^2}}{dt} &= p^2 \frac{1-\kappa}{2} U^a_{\boldsymbol{\eta}^2}
\end{aligned}
$$

Using this form, we end up with the following solution of the weight matrix

$$
w_i^a(t) = w_i^a(0) + \frac{\eta_i^1 U^a_{\boldsymbol{\eta}^1}(0)}{(1+\kappa)/2}\left[\exp\left(r^2 \frac{1+\kappa}{2} t\right) - 1\right] + \frac{\eta_i^2 U^a_{\boldsymbol{\eta}^2}(0)}{(1-\kappa)/2}\left[\exp\left(p^2 \frac{1-\kappa}{2} t\right) - 1\right] \quad (11)
$$

We therefore understand the following. At the beginning of the learning, since $r > p$, what is learned first is the mode toward the direction $\boldsymbol{\eta}^1 \propto \boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$, in a timescale that is given by time $t \sim 1/r^2$. At a different timescale, the part that is aligned with $\boldsymbol{\eta}^2$ will grow as well as discussed in the main text. Following the dynamics of the weights, as in eq. 11, we can infer the moment where the phase transition occurs. When considering an Hopfield like model, we know that the transition happens when

$$
\frac{\beta}{2N_{\rm v}}\left(\sum_i v_i \xi_i\right)^2 \sim \frac{1}{2N_{\rm v}}\left(\sum_i v_i \xi_i\right)^2
$$

that is, the critical temperature is $\beta_c = 1$. Following the dynamics of the weights of eq. 11, and neglecting the terms that are not aligned with $\boldsymbol{\eta}^1$ we can write

$$
\begin{aligned}
\frac{1}{2N_{\rm v}}\left(\sum_i v_i \xi_i\right)^2 &\sim \left(\frac{U^a_{\boldsymbol{\eta}^1}(0)}{(1+\kappa)/2}\right)^2 \left[\exp\left(r^2 \frac{1+\kappa}{2} t\right) - 1\right]^2 \frac{1}{2N_{\rm v}}\left(\sum_i v_i \eta_i^1\right)^2 \\
\beta_{\boldsymbol{\eta}^1}(t) &= \left(\frac{U^a_{\boldsymbol{\eta}^1}(0)}{(1+\kappa)/2}\right)^2 \left[\exp\left(r^2 \frac{1+\kappa}{2} t\right) - 1\right]^2
\end{aligned}
$$

where we can identify a sort of dynamical temperature associated to the pattern $\boldsymbol{\eta}^1$. Now, we need to be careful since by definition, the pattern $\boldsymbol{\eta}^1$ is made of random ± 1 components on its $(1+\kappa)/2$

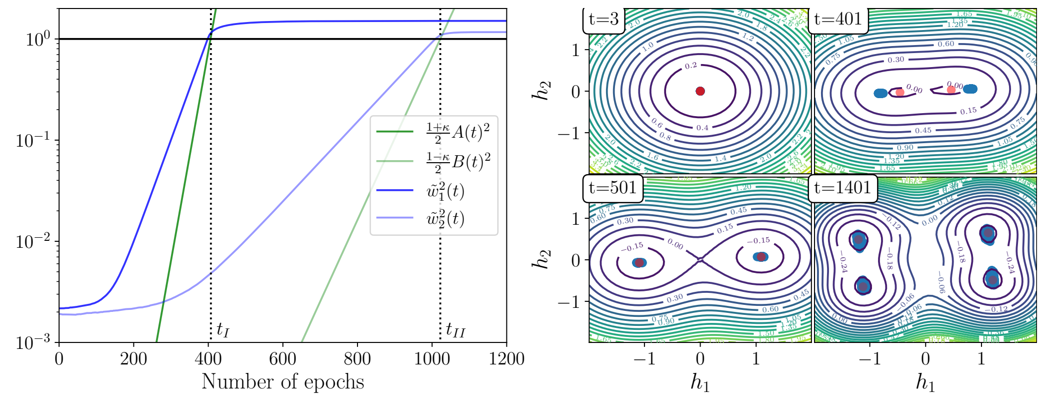

Figure 6: **Left:** the empirical dynamics of the eigenvalues of the weight matrix, here denotes $\tilde{w}_\alpha$ in blue. In green, the predicted dynamics as in eq. 11, adjusting only the initial conditions $U^a_{\boldsymbol{\eta}^1}(0)$ and $U^a_{\boldsymbol{\eta}^2}(0)$. We see that the curves cross the line $y = 1$ at the same moments t_I and t_{II}. **Right:** the free energy in the plane (h_1, h_2), the order parameters of the model. For different value of the weights during learning, we reconstruct the free energy of the system. We clearly see how the RBM first creates two minima, in the direction of $\boldsymbol{\eta}^1$, and then, split again to obtain the four fixed points.

elements and zero elsewhere. This rescales the critical temperature by a factor $(1 + \kappa)/2$. Therefore we need to look when

$$\frac{1+\kappa}{2}\beta_{\boldsymbol{\eta}^1}(t_I) \sim 1$$

and the same kind of argument can be used for the second transition with this time

$$\beta_{\boldsymbol{\eta}^2}(t) = \left(\frac{U^a_{\boldsymbol{\eta}^2}(0)}{(1-\kappa)/2}\right)^2 \left[\exp\left(p^2\frac{1-\kappa}{2}t\right) - 1\right]^2$$

$$\frac{1-\kappa}{2}\beta_{\boldsymbol{\eta}^2}(t_{II}) \sim 1$$

We show in Fig. 6, left panel how the times t_I and t_{II} compare with the moment where the eigenvalues w_α of the weight matrix cross the value one, which correspond to the phase transition following a statistical mechanics approach [8]. We observe that both indicators are crossing the line $y = 1$ at the same moment. In Fig. 6, right panel, we plot the behavior of the free energy (in the plane (h_1, h_2)). We see that at the moment of the transition, the free energy opens in the direction corresponding to the transition. Projecting the dataset (black dots) in the same direction as $h1$ (resp. $h2$), we can see how the system correctly positioned the minima once fully trained.

D The datasets and the rescaling

In this work, we illustrated our results on three datasets:

1. The Human Genome Dataset (HGD) [34] containing binary vectors, each representing a selection of 805 genes from a human individual, where 1s or 0s indicate the presence or absence of gene mutations relative to a reference sequence.
2. The MNIST dataset [42], containing 28×28 pixel black and white images of digitized handwritten digits.
3. The CelebA [43] dataset, in black-and-white, with 128×128 pixel images of celebrities faces.

The datasets MNIST and CELEBA, were either downscaled or upscaled in order to create dataset of various sizes. In practice, the function *resize* from the python library skimage was used either to increase or decrease the image size. The dataset HGD is geometrically a one-dimensional structure. In order to reduce its size, we took the convolution of each sample with a kernel of size $s = 3$. The output is one if the sum of the three input values (that are discrete variables in $\{0, 1\}$) of the kernel

dataset	N_{v}	N_{h}	ϵ	N_{ms}
MNIST	14×14	250	0.005	500
MNIST	19×19	500	0.005	500
MNIST	28×28	1000	0.005	500
MNIST	39×39	2000	0.005	500
MNIST	56×56	4000	0.001	500
CelebA	45×45	125	$4 \cdot 10^{-5}$	500
CelebA	64×64	250	$4 \cdot 10^{-5}$	500
CelebA	90×90	500	$4 \cdot 10^{-5}$	500
CelebA	128×128	1000	$4 \cdot 10^{-5}$	500
HGD	201	25	0.002	4500
HGD	402	50	0.002	4500
HGD	805	100	0.002	4500

Table 1: Hyperparameters of the RBMs analyzed in the main-text.

is above the threshold 2 and zero otherwise. A stride of 2 has been chosen such that the resulting samples has its size reduced by a factor two.

E Details on the training and the numerical analysis

E.1 Training

All RBMs analyzed in the main text were trained with the Persistent Contrastive Divergence (PCD) method [44] and $k = 100$ MCMC sampling steps per parameter update, to approximate the negative term of the log-likelihood gradient in Eq. (2). In this scheme, the last configurations reached in the MCMC process to compute the previous update are used as initialization of the new chains used to compute the subsequent update. This scheme tends to favor the equilibrium regime [29]. The results with other training schemes is discussed in section H. Moreover, as usual, we keep N_{chains} independent parallel Markov chains. For simplicity, N_{chains} is chosen to match the minibatch size used to estimate the positive term of the gradient. The code to reproduce the experiments is freely available in https://github.com/AurelienDecelle/TorchRBM. The hyperparameters used for each training (no. of visible and hidden units N_{v} and N_{h}, respectively, learning rate ϵ or minibatch size N_{ms}) are given in Table 1.

E.2 Numerical analysis

E.2.1 Susceptibility

Part of the analysis in section 5 of the main text is based on sampling the equilibrium configurations of models trained at different epochs to extract the moments of the distribution of magnetizations m_α, i.e. the projection of the samples along the different α-th singular vectors of $\boldsymbol{W}$. For this purpose, we automatically selected 10^3 models uniformly in logarithmic scale in training time and annealed the $N_{\mathrm{s}} = 1000$ independent samples from the least trained model to the most trained model, following the hot-to-cold thermal analogy, i.e. we perform N_{mesfr} alternate Gibbs sampling MCMC steps on each set of model parameters and use the last achieved configurations as a starting point to initialize the run on the next set of model parameters. The less trained model is initialized randomly. In parallel, we also consider the reverse scheme, where we consider a heating annealing. We start with the most trained machine, where all visible units are initialized to 1 (to force all initial configurations to be in a single cluster), and move backwards it in training time. We systematically checked that both analyses gave the same results for $N_{\mathrm{mesfr}} = 1000$ in the region of interest (the critical region) in Fig. 3 and 4.

In addition to the sampling procedure, we computed the SWD of the matrix $\boldsymbol{W}$ at each new parameter set and projected each visible configuration $\boldsymbol{v}$ along each of the left singular vectors $\boldsymbol{u}_\alpha$, to obtain $m_\alpha = (\boldsymbol{u}_\alpha \cdot \boldsymbol{v})/\sqrt{N_{\mathrm{v}}}$, and each hidden configuration $\boldsymbol{h}$ along the right singular $\bar{\boldsymbol{u}}_\alpha$ vector, to obtain $\bar{m}_\alpha = (\bar{\boldsymbol{u}}_\alpha \cdot \boldsymbol{v})/\sqrt{N_{\mathrm{v}}}$. The distribution moments are later estimated using the sample mean and variance of these N_{s} measures.

E.2.2 Hysteresis loop

To investigate the hysteresis behavior between different lumps at or above the phase transition, we first select the training epochs in which the first singular values of $\boldsymbol{W}$, w_1, take the values 4, 4.4 and 5. Note that for MNIST $w_{1,\mathrm{c}}$ is approximately 4. With the model parameters at each of these three selected number of updates, we perform N_s independent MCMC runs with the tilted Hamiltonian from Eq. (G). In these runs, the external field h is gradually varied to trace a loop: starting from $h = 0$, we slowly increase it to h_max, then decrease it to $-h_\mathrm{max}$ and finally bring it back to $h = 0$. Again, the last configurations reached at a certain h are used as initialization for the next one. In practice, we have chosen $h_\mathrm{max}/N_\mathrm{v}^{0.75}$, the increment in the field as $\delta h = 2 \times h_\mathrm{max}/N_\mathrm{loop}$ and $N_\mathrm{loop} = 50$. We can modulate the speed of the loop by performing a different number of MCMC steps k at each value of h. As shown in the figure, we consider $k =$ 10, 100, 1000 and 10^4. The results shown in the main text were obtained using the RBM trained with the original CELEBA dataset (i.e. $N_\mathrm{v} = 128 \times 128$), but the results obtained with other sizes and datasets are completely analogous.

F Link between the low-rank RBM and Mattis model

Let us consider a low-rank Ising-Ising RBM in which the $\mathcal{W}$ matrix has a single non-zero singular value ω, with left and right singular vectors $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$, and visible and hidden Ising variables (let's call these variables $\mathcal{W}$, $\boldsymbol{s}$ and $\boldsymbol{\tau}$ to distinguish them from the binary $0, 1$ version, which would be $\boldsymbol{W}$ and $\boldsymbol{v}$ and $\boldsymbol{h}$). In this case, the energy function of the RBM (if we ignore the biases for now) is

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\omega(\boldsymbol{u} \cdot \boldsymbol{v})(\bar{\boldsymbol{u}} \cdot \boldsymbol{h}),$$

which leads to a marginal energy on the visible

$$\mathcal{E}(\boldsymbol{s}) = -\sum_a \log\cosh\left[\sqrt{N_\mathrm{v}} w \bar{u}_a^2 m\right] \approx -\frac{1}{2} N_\mathrm{v} \omega^2 m^2 + O(m^4), \tag{12}$$

where we have defined $m = \boldsymbol{u} \cdot \boldsymbol{v}/\sqrt{N_\mathrm{v}}$ as the magnetization of the spins along the direction $\boldsymbol{u}$ and have exploited the fact that $\sum_a \bar{u}_a^2 = 1$ because it is a unit vector. One can obtain an analogous expression for the marginal energy on the hidden units, formulated in terms of the hidden magnetization $\bar{m} = \bar{\boldsymbol{u}} \cdot \boldsymbol{\tau}/\sqrt{N_\mathrm{h}}$. These energy functions, for small m or $\bar{m}$, are formally equal to those of the Mattis model for $\beta = \omega^2$, which means that our RBM should manifest a critical phase transition at $\beta_\mathrm{c} = T_\mathrm{c}^{-1} = \omega_\mathrm{c}^2 = 1$, with mean-field critical exponents. Standard RBMs are not formulated as Ising ± 1 variables, but in the form of binary $\{0, 1\}$ variables where we have the equivalence $4\boldsymbol{\mathcal{W}} = \boldsymbol{W}$ between the couplings matrices. This results in a critical point at $w_\mathrm{c} = 4$ and an effective inverse temperature $\beta = w^2/16$.

G Hysteresis in discontinuous transitions

The hysteresis phenomenon shown in Fig. 4-E is a classical measure in statistical physics and is a unique signature developed and observed in statistical physics to reveal that a high-dimensional probability measure had a phase transition where it splits into two distinct lumps. The procedure is to tilt the probability measure by introducing a contribution in the energy function that favors one lump over the other. In the present case, we use the learned preferred direction associated with the first phase transition $\boldsymbol{u}$, which is also the direction in which the probability measure splits, and we add a magnetic field to break the symmetry between the two modes created by the learning process. Therefore, we tilt the measure by adding

$$\mathcal{H}^\mathrm{tilted} = \mathcal{H}^\mathrm{RBM} - h \sum_i u_i v_i.$$

The tilted Hamiltonian favors one of the two lumps depending on the value of the local bias h. When the measure is actually concentrated on two distinct lumps, one of them leads to a sudden discontinuous transition ("first-order transition" in physics). In our case, the lumps are associated with the learned patterns and this additional contribution consists of the scalar product between the visible variables and the times of the learned patterns (the field that controls the strength of the tilting). In the presence of a first-order phase transition, one usually finds the phenomenon of hysteresis, i.e. the transition from one clump to another can be delayed due to metastability, leading to the characteristic

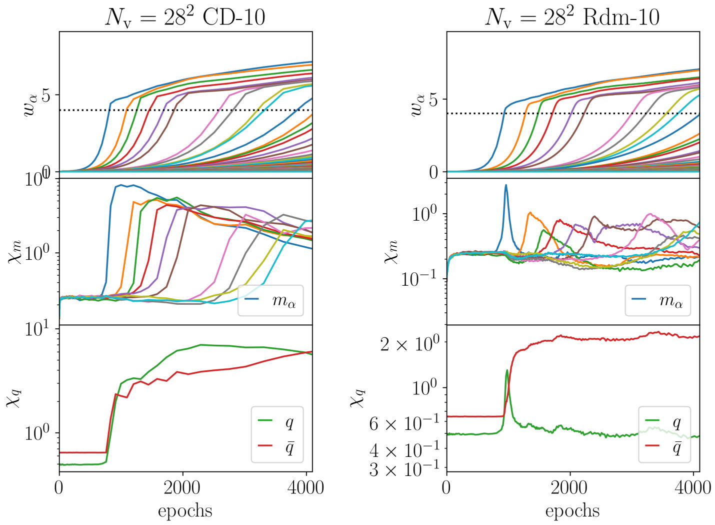

Figure 7: We reproduce Figs. 3A,B and C but for RBMs trained with CD-10 (left) and Rdm-10 (right).

hysteresis loops that we show in Fig. 4-E (see e.g. [45, 46] for a rigorous treatment and [47] for a physical treatment). This figure thus provides direct evidence for the decomposition of the measure into distinct lumps corresponding to the learned patterns, and that this decomposition occurs at the second-order phase transition that occurs during learning. The details about the numerical implementation are given in Section E.2.2.

H Results with other training schemes

All the machines analyzed in the main text were trained using the PCD-100 scheme, which involves initializing the chains with the PCD method and performing 100 MCMC steps per parameter update. This approach ensures that we obtain models in good equilibrium, avoiding the non-monotonic behavior in sample quality typical of out-of-equilibrium regimes [29]. However, it is more computationally expensive than standard methods, where only $k = 1 - 10$ steps are used, or alternative initialization strategies like Contrastive Divergence (CD) [48], where chains are initialized using the minibatch samples at each update, or the fully out-of-equilibrium regime (Rdm), where chains are always initialized randomly [29, 49].

In this appendix, we analyze RBMs trained with CD-10 and Rdm-10 strategies on the MNIST dataset. While the time evolution differs -often degrading susceptibility along learning directions— the overall picture of the cascade of phase transitions remains unchanged. We show the equivalent to Figs. 3A,B,C with these two new trainings in Fig. 7.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: .

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: .

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The theoretical analysis relies on theoretical physics development. In such context, the full set of assumptions is not state clearly, but the results is expected to be correct within the range of application. As such no theorem or lemma is provided.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The method used here is quite standard (RBM training in a simple case), all the rest is clearly specified.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The experiment perform are sufficiently precise within the claim of the paper so they do not need further error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: the experiments do not need particular resources and can be trained on personal laptop.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: .

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: .

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: .

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: .

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: .

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: .

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: .

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.