


‘Place-cell’ emergence and learning of invariant data with restricted Boltzmann machines: breaking and dynamical restoration of continuous symmetries in the weight space

Moshir Harsh¹, Jérôme Tubiana², Simona Cocco¹
and Rémi Monasson^{1,3} 

¹ LPENS, Ecole Normale Supérieure, CNRS UMR 8023 and PSL Research, 24 Rue Lhomond, 75231 Paris Cedex 05, France

² Blavatnik School of Computer Science, Tel Aviv University, Israel

Received 29 December 2019, revised 3 March 2020

Accepted for publication 5 March 2020

Published 8 April 2020



CrossMark

Abstract

Distributions of data or sensory stimuli often enjoy underlying invariances. How and to what extent those symmetries are captured by unsupervised learning methods is a relevant question in machine learning and in computational neuroscience. We study here, through a combination of numerical and analytical tools, the learning dynamics of restricted Boltzmann machines (RBM), a neural network paradigm for representation learning. As learning proceeds from a random configuration of the network weights, we show the existence of, and characterize a symmetry-breaking phenomenon, in which the latent variables acquire receptive fields focusing on limited parts of the invariant manifold supporting the data. The symmetry is restored at large learning times through the diffusion of the receptive field over the invariant manifold; hence, the RBM effectively spans a continuous attractor in the space of network weights. This symmetry-breaking phenomenon takes place only if the amount of data available for training exceeds some critical value, depending on the network size and the intensity of symmetry-induced correlations in the data; below this ‘retarded-learning’ threshold, the network weights are essentially noisy and overfit the data.

Keywords: restricted Boltzmann machine, machine learning, receptive field, learning

(Some figures may appear in colour only in the online journal)

³ Author to whom any correspondence should be addressed.

1. Introduction

Many high-dimensional inputs or data enjoy various kinds of low-dimensional invariances, which are at the basis of the so-called manifold hypothesis [1]. For instance, the pictures of somebody's face are related to each other through a set of continuous symmetries corresponding to the degrees of freedom characterizing the relative position of the camera (rotations, translations, changes of scales) as well as the internal deformations of the face (controlled by muscles). While well-understood symmetries can be explicitly taken care of through adequate procedures, e.g. convolutional networks, not all invariances may be known *a priori*. An interesting question is therefore if and how these residual symmetries affect the representations of the data achieved by learning models.

This question does not arise solely in the context of machine learning, but is also of interest in computational neuroscience, where it is of crucial importance to understand how the statistical structure of input stimuli, be they visual, olfactory, auditory, tactile, ... shapes their encoding by sensory brain areas and their processing by higher cortical regions. Information theory provides a mathematical framework to answer this question [2], and was applied, in the case of linear models of neurons, to a variety of situations, including the prediction of the receptive fields of retinal ganglion cells [3], the determination of cone fractions in the human retina [4] or the efficient representation of odor-variable environments [5]. In the case of natural images, which enjoy approximate translational and rotational invariances, non-linear learning rules resulting from adequate modification of Oja's dynamics [6] or sparse-representation learning procedures [7] produce local edge detectors, such as independent component analysis [8]. These detectors bear strong similarities with the neural receptive fields measured in the visual cortex (V1 area) in mammals.

It is therefore natural to wonder whether the existence of localized receptive fields is a general feature to be expected from representations of invariant distributions of inputs. Gardner's theory of optimal learning for single-layer neural networks (perceptron) [9] predicts that spatially correlated patterns, e.g. drawn from a translationally-invariant distribution, lead to a localized profile of weights [10]. Further supporting evidence was recently brought by several works, focusing on the production of such receptive fields in the context of unsupervised learning. Learning of symmetric data with similarity-preserving representations [11] or with auto-encoders [12] both led to localized receptive fields tiling the underlying manifold, in striking analogy with place cells and spatial maps in the hippocampus. In turn, such high-dimensional place-cell-like representations have putative functional advantages: they can be efficiently and accurately learned by recurrent neural networks, and thus allow for the storage and retrieval of multiple cognitive low-dimensional maps [13].

The present work is an additional effort to investigate this issue in a highly simplified and idealized framework of unsupervised learning, where both the data distribution and the machine are under full control. Similarly to previous studies [14, 15], we consider synthetic data with controlled invariances generated by standard models in statistical physics, such as the Ising and XY models. These data are then used to train restricted Boltzmann machines (RBM), a simple albeit powerful framework for representation learning, where a layer of hidden (latent) units account for the correlation structure in the data configurations. We show how the receptive fields of the hidden units undergo a symmetry-breaking transition in the space of couplings: units individually cover localized regions of the input space, but concur to tile the space as best as possible, in much the same way as hippocampal place cells do. Translation invariance, present in the data distribution but broken by each hidden unit, is dynamically restored if we let the training algorithm run for very long times (well beyond the training time needed to saturate the log-likelihood of the test set): while keeping their localized shape, the center of the

receptive/place fields diffuses along the input space, effectively ensuring the invariance of the learned distribution. We also show that this symmetry-breaking phenomenon requires a minimum number of data, an illustration of the general phenomenon of retarded learning [16], also encountered in random matrix theory in the context of the so-called spiked covariance model [17, 18].

Our paper is organized as follows. RBM and their learning algorithms are introduced in section 2. We consider the case of a data distribution with a single invariance in section 3, and with two symmetries in section 4. A detailed theoretical analysis of the learning dynamics and of the receptive field emerging through the symmetry-breaking transition can be found in section 5. Conclusions and speculative connections with experiments in neuroscience are proposed in section 6.

2. Restricted Boltzmann Machines

2.1. Definition and log-likelihood

A restricted Boltzmann machine (RBM) is a bipartite, undirected stochastic neural network with two layers, see figure 1:

- the *visible* layer includes N units v_i , $i = 1, \dots, N$, which carry the configurations of data. For simplicity, we assume here that visible units take binary values, $v_i = \pm 1$.
- the *hidden* layer includes M units h_μ , $\mu = 1, \dots, M$, on which are expressed the representations of the data configurations. Hidden, or latent variables h_μ can take real or binary values.

The model is formally defined by a Gibbs probability distribution over the sets of visible (\mathbf{v}) and hidden (\mathbf{h}) variable configuration:

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}} e^{-E(\mathbf{v}, \mathbf{h})}, \quad \text{where } \mathcal{Z} = \sum_{\mathbf{v}} \int d\mathbf{h} e^{-E(\mathbf{v}, \mathbf{h})} \quad (1)$$

is the partition function, such that p is normalized to unity, and the energy function $E(\mathbf{v}, \mathbf{h})$ is given by

$$E(\mathbf{v}, \mathbf{h}) = - \sum_{i=1}^N \sum_{\mu=1}^M w_{i\mu} v_i h_\mu - \sum_{i=1}^N b_i v_i + \sum_{\mu=1}^M \mathcal{U}_\mu(h_\mu). \quad (2)$$

In the formula above, $w_{i\mu}$ are the real-valued weights (coupling) connecting the hidden unit h_μ and the visible unit v_i , b_i are real-valued bias terms, also called fields and \mathcal{U}_μ are the hidden unit potentials. We consider two possible choices for \mathcal{U}_μ :

- For binary (± 1) valued hidden units, a regular field term $\mathcal{U}_\mu(h_\mu) = -c_\mu h_\mu$ similar to the visible units. In that case, equation (2) is a special case of Ising distribution, with only couplings between units belonging to different layers.
- For real valued hidden units, the symmetric double well potential $\mathcal{U}_\mu(h_\mu) = \frac{1}{2} h_\mu^2 + \theta_\mu |h_\mu|$. For $\theta_\mu = 0$, the potential is quadratic and the corresponding variable is Gaussian and for $\theta_\mu < 0$ the potential has two minimas at $\pm \theta_\mu$; this choice of potential effectively interpolates between Gaussian ($\theta_\mu = 0$) and binary ($\theta_\mu \rightarrow -\infty$) hidden units [19]. These hidden units are popularly referred to as the **rectified linear units (ReLU)** with the difference that here we have made the potential symmetric for positive and negative values of the h_μ .

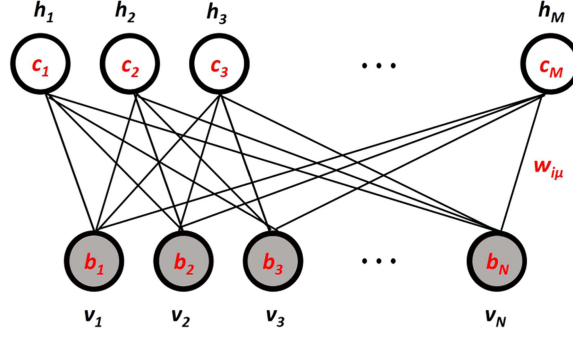


Figure 1. The two-layer structure of RBM, with weights $w_{i\mu}$ connecting N visible units v_i to M hidden units h_μ . These binary-valued units are subjected to local fields, called b_i and c_μ for, respectively, the visible and hidden layers.

Due to the absence of connections between the units within a layer, the conditional probability of hidden units given the visible units factorizes as follows:

$$p(\mathbf{h}|\mathbf{v}) = \prod_{\mu=1}^M p(h_\mu | I_\mu(\mathbf{v})), \quad (3)$$

where $I_\mu(\mathbf{v}) = \sum_i w_{i\mu} v_i$ is the total input received from the visible layer by hidden unit μ in the absence of fields on visible units, and $p(h_\mu | I) \propto e^{\mathcal{U}_\mu(h_\mu) + h_\mu I}$. Therefore, sampling from the conditional distribution is simply done by first computing the hidden layer inputs I_μ , then sampling independently each hidden unit given its input according to its hidden unit potential. Similarly, the average activity of a hidden unit given the visible units, $\langle h_\mu | \mathbf{v} \rangle$, is a non-linear function of the input $I_\mu(\mathbf{v})$; for binary hidden units, we have $\langle h_\mu | \mathbf{v} \rangle = \tanh(\sum_i w_{i\mu} v_i + c_\mu)$. Therefore, RBM can be viewed as a nonlinear model similar to other feature extraction methods such as independent component analysis. Symmetric formulas can be written for the conditional probability of visible units given the hidden units.

In addition, the marginal distribution over the visible units $p(\mathbf{v})$ can be written in closed form:

$$\begin{aligned} p(\mathbf{v}) &= \int d\mathbf{h} p(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}} e^{\sum_{i=1}^N b_i v_i} \prod_{\mu=1}^M \int dh_\mu e^{-\mathcal{U}_\mu(h_\mu) + h_\mu I_\mu(\mathbf{v})} \\ &= \frac{1}{\mathcal{Z}} \exp \left(\underbrace{\sum_{i=1}^N b_i v_i + \sum_{\mu=1}^M \Gamma_\mu(I_\mu(\mathbf{v}))}_{-E_{\text{eff}}(\mathbf{v})} \right), \end{aligned}$$

where $\Gamma_\mu(I) = \log \int dh e^{-\mathcal{U}_\mu(h) + h I}$ is the cumulant generative function, or log Laplace transform, associated to the potential \mathcal{U}_μ ; for binary hidden units, $\Gamma_\mu(I) = \log 2 \cosh(I + c_\mu)$. Note that by construction, $\Gamma'_\mu(I_\mu)$ is the average value of the hidden unit given its input I_μ ; therefore the hidden unit potential determines the transfer function of the hidden unit. Importantly, although the joint distribution is pairwise, the marginal distribution is not in general as Γ_μ functions are not quadratic. Therefore, RBM generate effective high-order interactions between

the units v_i , and are capable of expressing complex measures over the visible configurations [20, 21].

2.2. Training algorithm

Training the RBM is the process of fitting the parameters $\Theta = \{w_{i\mu}, b_i, c_\mu/\theta_\mu\}$ to maximize the average log-likelihood of the S data items \mathbf{v}^{data} assumed to be independently drawn from $p(\mathbf{v})$. While this may be done with the gradient ascent method, calculating the likelihood is computationally intensive as it requires evaluating the partition function, and sampling methods like Markov chain Monte Carlo (MCMC) in the form of Gibbs sampling are used.

2.2.1. Gradient of log-likelihood. For the model with parameters Θ , the log-likelihood of a single training example \mathbf{v}^{data} is

$$\log \mathcal{L}(\mathbf{v}^{\text{data}}|\Theta) = \log p(\mathbf{v}^{\text{data}}) = -E_{\text{eff}}(\mathbf{v}^{\text{data}}) - \log \mathcal{Z} = -E_{\text{eff}}(\mathbf{v}^{\text{data}}) - \log \left[\sum_{\mathbf{v}} e^{-E_{\text{eff}}(\mathbf{v})} \right]. \quad (4)$$

Taking the partial derivative with respect to Θ gives

$$\frac{\partial \log \mathcal{L}(\mathbf{v}^{\text{data}}|\Theta)}{\partial \Theta} = -\frac{\partial E_{\text{eff}}(\mathbf{v}^{\text{data}})}{\partial \Theta} + \frac{1}{\mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial \Theta} = -\frac{\partial E_{\text{eff}}(\mathbf{v}^{\text{data}})}{\partial \Theta} + \left\langle \frac{\partial E_{\text{eff}}(\mathbf{v})}{\partial \Theta} \right\rangle_{\text{RBM}}, \quad (5)$$

where $\langle (\cdot) \rangle_{\text{RBM}} = \frac{1}{\mathcal{Z}} \sum_{\mathbf{v}} e^{-E_{\text{eff}}(\mathbf{v})} (\cdot)$ denotes the average according to the marginal distribution over the visible units with parameter values Θ .

In particular, for the weights $w_{i\mu}$, we have according to (5), $\frac{\partial E_{\text{eff}}(\mathbf{v})}{\partial w_{i\mu}} = -v_i \Gamma'_\mu(I_\mu(\mathbf{v})) \equiv -v_i \langle h_\mu | \mathbf{v} \rangle$. The gradient of the total log-likelihood is then

$$\frac{\partial \langle \log \mathcal{L}(\mathbf{v}^{\text{data}}|\Theta) \rangle_{\text{data}}}{\partial w_{i\mu}} = \langle v_i^{\text{data}} \langle h_\mu | \mathbf{v}^{\text{data}} \rangle \rangle_{\text{data}} - \langle v_i \langle h_\mu | \mathbf{v} \rangle \rangle_{\text{RBM}}. \quad (6)$$

Equation (6) is an example of a moment-matching condition, as it imposes that the correlation between the variables v_i and h_μ computed from the data coincides with its counterpart defined by the RBM model distribution $p(\mathbf{v}, \mathbf{h})$. The gradients of \mathcal{L} over b_i and c_μ lead to similar moment-matching conditions for, respectively, the average values of v_i and of h_μ .

2.2.2. Approximating the log-likelihood gradient. In the gradient of the log-likelihood of equation (6), the model-distribution moment is not computationally tractable, as it requires to sum over all values of the visible and the hidden variables. In practice, an approximate value for this term is obtained by Markov chain Monte Carlo (MCMC) methods. The Markov Chain is defined by repeated iterations of Gibbs sampling, which consists in sampling \mathbf{h} from \mathbf{v} and \mathbf{v} from \mathbf{h} using equation (3). In principle, one should run a full MCMC simulation at each gradient step, but this is computationally prohibitive. For our RBM training we use the *persistent contrastive divergence* (PCD) algorithm [22]: Markov chains are initialized at the beginning of the training and updated with only a few Gibbs Monte Carlo steps between each evaluation of the gradient, see [23] for a more detailed review. This approximation works very well for the data distribution studied here because they are in a paramagnetic phase (=monomodal), hence the Markov chains mix very rapidly.

2.2.3. Stochastic optimization. The RBM is trained using stochastic gradient ascent (SGA), the golden standard for neural network optimization. SGA is a variant of ordinary gradient ascent where at each step, only a small subset of the data set (the minibatch), of size $S_{\text{batch}} \sim 10\text{--}100$ examples, is used to evaluate the average log-likelihood, see equation (7) where ν is the learning rate which dictates how much to change the parameter in the direction of the steepest gradient. The dataset is divided into S/S_{batch} mini-batches $\text{batch}(t)$, and for each epoch of training t , we perform one SGA update for each mini-batch. An epoch consists of using all the subsets for the update such that each data sample is used once. After every epoch the subsets are again drawn randomly. Several dozens of epochs are usually required to reach convergence.

$$\Theta^{t+1} = \Theta^t + \nu \left[\frac{1}{S_{\text{batch}}} \sum_{b \in \text{batch}(t)} \nabla_{\Theta} \log \mathcal{L}(\mathbf{v}^{\text{data},b} | \Theta) \right] \quad (7)$$

Compared to ordinary gradient ascent, SGA serves several purposes. First and foremost, its computational cost is significantly reduced as only a small batch is used per update; yet the update is usually reliable thanks to data redundancy. Second, the stochastic evaluation of the gradient introduces noise in the learning process. This prevents the dynamics from getting trapped in local maxima, which is crucial for non-convex optimization landscapes, and it also directs the dynamics toward minima with wider basins of attraction [24]. It has been argued that the later effect contributes in improving generalization performance [25–27]. Though the convergence rate of SGA has a slower asymptotic rate than ordinary gradient descent, it often does not matter in practice for finite data sets, as the performance on the test set usually does not improve anymore once the asymptotic regime is reached [28].

The noise level of the SGD is directly related to the batch size and learning rates parameters, see for instance [29]. Briefly speaking, assuming i.i.d. and infinite number of samples, the SGA parameter increment has mean value $\nu \nabla_{\Theta} \langle \log \mathcal{L}(\mathbf{v}^{\text{data}} | \Theta) \rangle_{\text{data}}$, and variance proportional to ν^2/S_{batch} ; in the large S_{batch} limit it is also Gaussian distributed according to the central limit theorem. In comparison, the increments of a continuous time Langevin equation with energy landscape $E = -\langle \log \mathcal{L}(\mathbf{v}^{\text{data}} | \Theta) \rangle_{\text{data}}$ and noise covariance matrix $\propto \sigma_{\text{SGA}}^2$, integrated over a time step ν has the same mean value and a covariance proportional to $\sigma_{\text{SGA}}^2 \nu$. Identifying both noises gives the following scaling law for the SGA noise, $\sigma_{\text{SGA}} \propto \sqrt{\frac{\nu}{S_{\text{batch}}}}$. Reducing the learning rate and increasing the batch size therefore decrease the noise level, and vice-versa. In all our experiments, both learning rates and batch sizes are kept fixed throughout a training session.

3. Learning data with a single invariance

3.1. Data distribution and translation invariance

3.1.1. Ising model. Our first toy distribution for data \mathbf{v}^{data} is the celebrated one-dimensional Ising model from statistical physics. Here each v_i is a spin which can either be up or down, that is can take only ± 1 binary values. The corresponding joint probability distribution of the visible units reads

$$p_{\text{data}}(v_1, v_2, \dots, v_N) = \frac{1}{Z_{\text{Ising}}} e^{\beta \sum_{i=1}^N v_i v_{i+1}} \quad (8)$$

where the partition function normalizes this probability over the 2^N visible configurations, and $\beta > 0$ is referred to as the inverse temperature. We enforce periodic boundary conditions through $v_{N+1} \equiv v_1$.

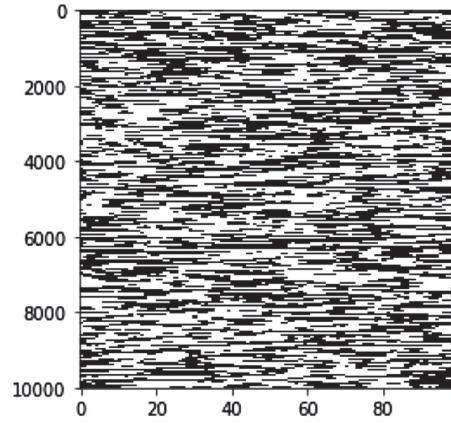


Figure 2. 10 000 one-dimensional Ising model configurations with 100 spins each, sampled from distribution (8) at inverse temperature $\beta = 1$ and with periodic boundary conditions. Black and white dots represent units equal to, respectively, $+1$ and -1 . The correlation length ξ may be thought of as the typical length of black or white contiguous regions along the horizontal direction. Here, $\beta = 1$, which corresponds to $\xi \simeq 3.7$.

As is well known, under distribution (8), all visible units v_i have average values equal to zero, and the correlation function decays exponentially with the distance separating the corresponding units on the ring,

$$\langle v_i v_j \rangle = \sum_{\mathbf{v}} p_{\text{data}}(v_1, v_2, \dots, v_N) v_i v_j = e^{-|i-j|/\xi}, \quad \text{where } \xi = \frac{1}{\ln \coth \beta} \quad (9)$$

is the correlation length. The above expression for the correlation holds when $N \gg \xi$.

Formula (8) defines a simple example of invariant distribution under the set of translations (or, better, rotations) on the N -site ring. More precisely, for any integer k , we have

$$p_{\text{data}}(v_1, v_2, \dots, v_N) = p_{\text{data}}(v_{k+1}, v_{k+2}, \dots, v_{k+N}), \quad (10)$$

where $i + k$ is to be intended modulo N . Figure 2 shows a number of configurations, drawn independently and at random from this probability distribution using the Gibbs sampling algorithm.

3.1.2. Symmetry breaking and weight profile. The usual meaning of symmetry breaking in statistical mechanics, condensed matter or field theory is that the expectation value of some observable (typically, the field over space, i.e. magnetization) becomes non zero below some critical temperature. In the context of weight learning by RBM, this would correspond to the fact that the average value of the weights, e.g. attached to a hidden unit,

$$W_\mu = \sum_i w_{i\mu}, \quad (11)$$

would become non zero for well chosen control parameter, e.g. above some critical value. Due to the translation invariance in the data, the simplest, symmetric scenario corresponding to $W_\mu \neq 0$ would be that all weights are equal to $w_{i\mu} = W/N$.

In the following, we will refer to symmetry breaking for situations in which both $W_\mu \neq 0$ and $w_{i\mu} \neq W/N$. In other words, symmetry breaking hereafter refers to the emergence of RBM

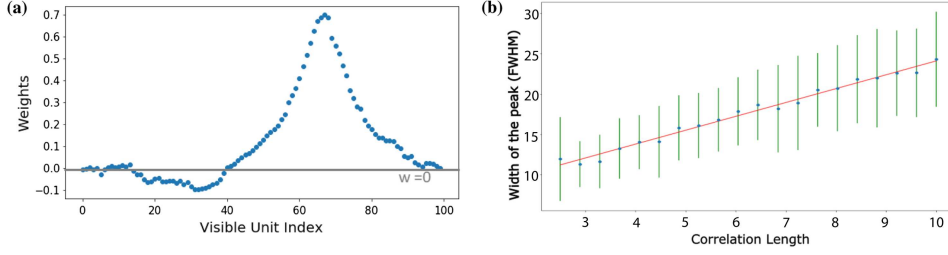


Figure 3. (a) Structure of weights learned by a RBM trained weights after 100 epochs from 10 000 data configurations of the one-dimensional Ising model of size $N = 100$. Training parameters: rate $\nu = 0.001$, batch size $S_{\text{batch}} = 10$, PCD-20 learning. One observes the emergence of a peaked structure in the weights, centered around site $i \simeq 68$. Note the small fluctuations in the tails (small w_{i1}), due to the finite (but large) number of data. (b) Width of the peak in the weight space as a function of the correlation length of the Ising model, ξ . We trained our RBM with one hidden unit 25 times on data generated at different temperatures, β , and then calculated the average peak width and the standard deviation (error bars) over the different samples. The width was calculated by fitting a cubic spline with one knot to the profile: $y(i) = \{w_i - \max(w_i)/2\}$, where i is the site index. The roots of this spline were then determined numerically, and the width was defined as the modulus of the difference between the roots. This procedure reliably finds the full width at half maximum (FWHM). A linear fit (red line) of the form $y = ax + b$ shows that the width of the place (receptive) field of the only hidden unit is proportional to the relevant characteristic length in the data. Notice that the intercept (b) is non zero, in agreement with the theoretical findings of section 5.2 in the $\beta \rightarrow 0$ limit.

model distributions (defined through the weights $w_{i\mu}$) that do not enjoy the data symmetries, such as translation invariance in the case of the Ising data in (10). The corresponding weight profile will therefore vary over the input (visible) space, and will be called spatially structured.

3.2. Initial learning and emergence of place cells

3.2.1. Case of a single hidden unit. First we train the RBM with only $M = 1$ hidden unit, and $N = 100$ visible units. Such a limited machine is, of course, not expected to reproduce accurately the Ising model distribution underlying the data. However, this is an interesting limit case to study how the RBM can make the most of its single set of weight attached to the unit. We use a large number of data configurations for training, which makes our distribution approximately invariant under rotations on the ring.

We initialize the weights w_{i1} with small amplitude Gaussian random values; since the data are symmetric, we further impose $b_i = c_\mu = 0 \forall i, \mu$. We observe that the log-likelihoods of the test and training sets saturate after about 10–20 epochs of training. The results of the training after 100 epochs, i.e. much after saturation is reached are shown in figure 3(a). We observe that the weights w_{i1} are not uniform as could have been naively expected from rotational invariance, but focus on a limited portion (place) of the N -site ring. The position of the peak depends on the initial conditions for the weights; it may also be influenced by the small irregularities in the data set coming from the finite number of training configurations.

To understand what determines the width of the weight peak, we train different RBMs with data at different inverse temperatures β , and calculate their average peak widths over multiple runs. We plot the peak width as a function of the correlation length ξ in figure 3(b). We observe

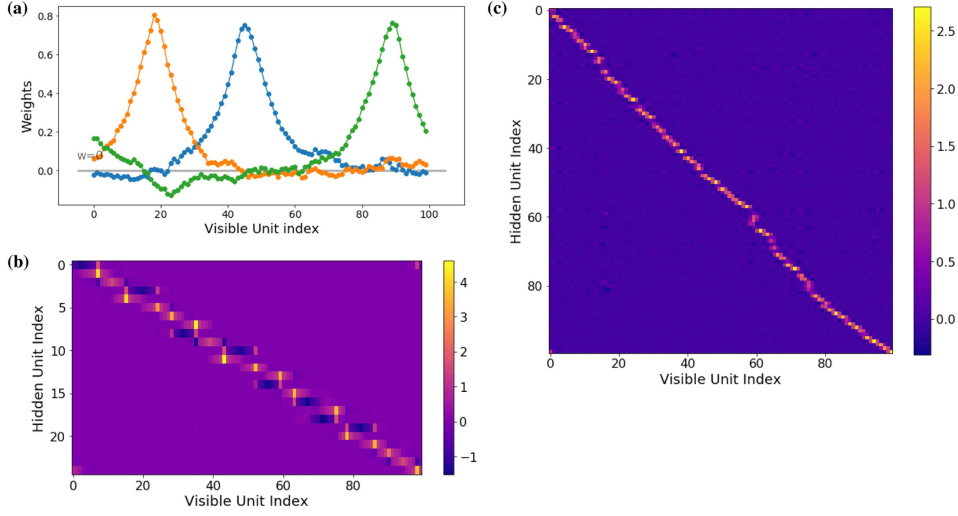


Figure 4. (a) Same as figure 3(a), but with a RBM having $M = 3$ hidden units. The weights attached to the same hidden units are shown with the same color. The receptive fields (peaks) for the three hidden units are roughly equally separated from each other. Note that due to the invariance of the probability distributions of the RBM under $h_\mu \rightarrow -h_\mu, w_{i\mu} \rightarrow -w_{i\mu}$ (when $c_\mu = 0$), the overall sign of the weights attached to the same hidden unit does not matter. (b) and (c) Same as figure 3(a), but with a RBM having, respectively $M = 25$ (b) and $M = 100$ (c) hidden units. The color codes show the intensity of the weights $w_{i\mu}$ as a function of the hidden (μ , y-axis) and visible (i , x-axis) unit labels. The hidden units have been arranged according to the center of their respective receptive field.

that the peak width scales proportionally to ξ .³ Interestingly, despite its very limited expression power, our single-unit RBM has correctly learned to coarse grain the visible unit configurations on the relevant scale length in the data, ξ . Having wider receptive, or place fields would not be as much as informative. For instance, with a set of uniform weights $w_{i1} = w$, the hidden unit would simply estimate the average magnetization of (mean value of all visible units in) the data configurations, which are all equal to zero up to fluctuations of the order of $\pm N^{-1/2}$, and would completely miss the correlated structure of the data. Conversely, more narrow place fields would have lower signal-to-noise ratios: the strong correlations of visible units over the length ξ allows one to reliably estimate the local magnetization and the correlation structure on this scale.

3.2.2. Case of multiple hidden units. We next show results obtained when training RBM with $M = 3$ hidden units on the same data. Figure 4 shows that each one of the three sets of weights have roughly the same peaked structure (same width) as in the $M = 1$ case, but the peaks are centered at different places along the ring. The roughly equal distance between successive peaks shows the existence of an effective repulsion between the weights of any two hidden units. This phenomenon is easy to understand on intuitive grounds: having very overlapping place fields produces highly redundant hidden units, and would not help capturing the spatial correlation in the data spreading over the entire ring.

⁴ We have checked that the width is independent of the size N for sufficiently large values of N .

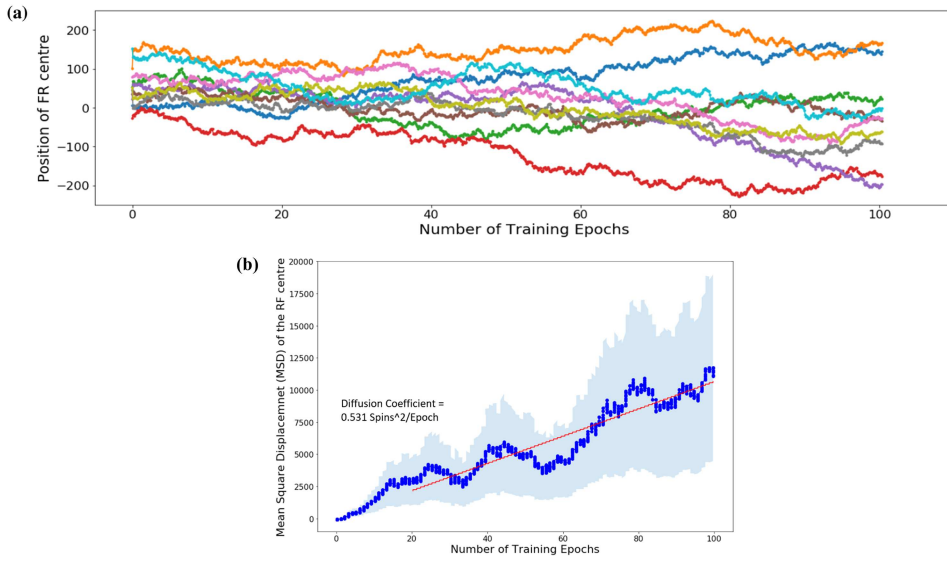


Figure 5. (a) Trajectories of the peak of the receptive field for a RBM with one hidden unit vs number of epochs of training. Each trajectory corresponds to a run of the training procedure starting from a random initial condition for the weights. Training parameters: rate $\nu = 0.1$, PCD-20 training and batch size $S_{\text{batch}} = 100$, 1 million configurations of the one-dimensional Ising model at $\beta = 1$. Trajectories are corrected for periodic boundary conditions—when the peak crosses the boundary, we add or subtract the size N of the system. (b) Mean square displacement (MSD) of the peak of the receptive field vs number t of epochs of training. Time $t = 0$ corresponds to the beginning of training. The red line is a linear fit $= 2Dt$ obtained after removing the transitory period of the first 20 epochs of training. Results obtained from the trajectories shown in (a).

Training of RBMs with a large number of hidden units shows the same pattern of production of place fields attached to different hidden units, covering in an approximately uniform way the visible space (ring), see figure 4(c) in the case of $M = 100$ hidden units. The only notable difference is that the width of the place fields shrinks as M gets very large. This happens when $M\xi \gg N$, i.e. when the single-hidden-unit peaks would start to largely overlap.

3.3. Long-time learning and restoration of invariance through place-field diffusion

We now let the training dynamics evolve for a much larger number of epochs. In the case of a RBM with one hidden unit only, the weight vector shows the overall peak structure of figure 3(a) at all times (after a short initial transient of about 10 epochs during which the localized peak emerges). However, the location of the peak may change on very long training time scales. Figure 5(a) shows ten trajectories of the center of the peak corresponding to ten random initializations of the weights (equal to small values drawn independently and randomly). We observe that the centers of the peaks undergo a seemingly random motion. When the number of data items used for training is very large (to erase any tiny non-homogeneity in the empirical distribution), this random motion looks like pure diffusion.

Figure 5(b) shows that the mean square displacement (MSD) of the peak center grows roughly linearly with the training time (number of epochs), which defines the effective diffusion coefficient of the weight peak. For intermediate numbers of data items, diffusion is activated: due to the inhomogeneities in the empirical data distribution, some places along the ring are preferred, and have a tendency to trap the weight peak for some time.

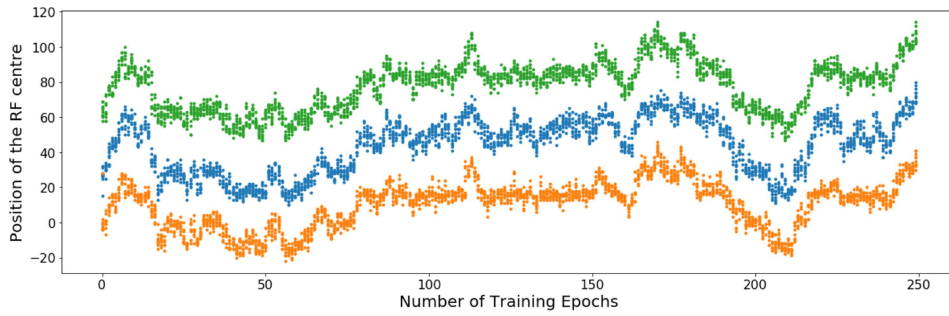


Figure 6. Locations of the peaks of the receptive fields for a RBM with $M = 3$ hidden units as functions of the number of epochs of training. Same training parameters as in figure 5. The trajectories are corrected for periodic boundary conditions—when the peak crosses the boundary, we add or subtract the size N of the system. Multiple trajectories are repeated runs of the training procedure, starting from different initial random conditions for the weights. The correlated motion of the place fields is a clear signature of the presence of repulsive interactions between the corresponding weight vectors.

Repeating the same analysis for a RBM with $M = 3$ hidden units allows us to observe the diffusion of the three peak centers, see figure 6. We see that the motions of these centers are coupled to maintain a constant distance between each other. This is a clear signature of the effective repulsion between the hidden-unit weight vectors already discussed in section 3.2.2.

3.4. Case of few data: retarded learning transition

The emergence of a pronounced peak in the weight vector attached to a hidden unit reported above takes place only if the number of data items are sufficiently large. For very few data, the RBM weights do not show any clear receptive field and seem to overfit the data. Similarly, for a fixed number of data samples, a transition is observed between the unstructured and spatially-structured regimes as the correlation length ξ (or the inverse temperature β), that is, the spatial signal in the data is increased. To distinguish these two regimes, we introduce the empirical order parameter, see definition (11),

$$W = \left| \sum_{i=1}^N w_{i1} \right|, \quad (12)$$

which is expected to be large when place fields emerge, and much smaller (and vanishingly small in the large- N limit) in the unstructured regime. Note that the definition above for W relies on the positive nature of the correlations between the visible units; in the case of anti-correlations, staggered versions of W should be considered instead.

Figure 7(a) shows the value of the order parameter W as a function of the intensity of spatial correlations for a fixed number of data samples. For small values of β (and ξ) W vanishes: the very weak spatial structure in the available data is not learned by the RBM. At large β , a place field emerges, focusing on a finite portion of the ring, and W is non zero. The same transition is observed when β is fixed and the number of training samples, S , is varied, see figure 7(b). For few samples or, equivalently, large noise levels $r = N/S$, the RBM overfits the data in the sense that no spatial structure is learned, and W vanishes. For small values of r , W becomes non zero, signaling the emergence of a place field focusing on a finite portion of the ring.

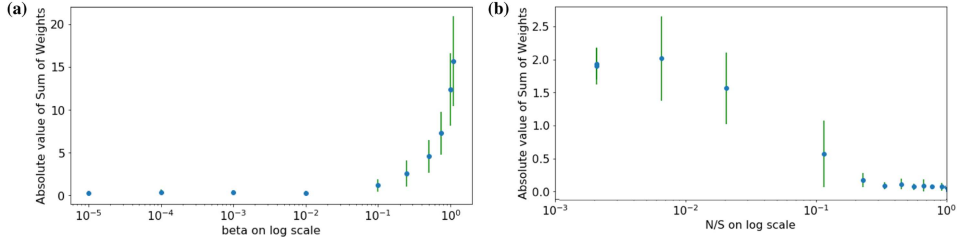


Figure 7. (a) Sum of weights W after training a one-hidden-unit RBM on 10 000 configurations of the one-dimensional Ising model at different β . Here, there are $N = 100$ visible units, i.e. noise ratio $r = N/S = 0.01$, other RBM training parameters are: $\nu = 0.001$, $S_{\text{batch}} = 10$, trained for 250 epochs with PCD-20. The error bars are calculated over 10 runs of the same training. (b) Sum of weights W after training a one-hidden-unit RBM on configurations of the one-dimensional Ising model at $\beta = 0.5$ as a function of the noise ratio $r = N/S$. Here, there are $N = 20$ visible units. All other parameters of training are same as (a). The error bars are calculated over 10 runs of the same training.

This transition is an example of the very general mechanism of the so-called retarded learning phenomenon [16], also encountered in the context of random correlation matrices and the spiked covariance model. The connection with random matrices will be made explicit in section 5.1.

4. Learning data with multiple invariances

4.1. Data distribution: discretized XY model

The classical XY model is a popular model in statistical physics, used in particular to study topological phase transitions in two dimensions. We consider here the one-dimensional version of this model, which shows no such phase transition but is nonetheless very useful for our study due to the additional symmetry with respect to the Ising model. In the XY model each lattice site i carries an angle $\theta_i \in [0, 2\pi[$ with respect to some arbitrary, fixed direction. The energy function reads, up to a scale factor that can be absorbed in the temperature definition,

$$E(\theta_1, \theta_2, \dots, \theta_N) = - \sum_{i=1}^N \cos(\theta_i - \theta_{i+1}) \quad (13)$$

with periodic boundary condition $\theta_{N+1} = \theta_1$. We then discretize the set of angle values in multiples of $2\pi/P$, where P is an integer. The resulting model is a Potts model over the N integer-valued variables $v_i = 0, 1, 2, \dots, P-1$, with probability distribution (with periodic boundary conditions)

$$p_{\text{data}}(v_1, v_2, \dots, v_N) = \frac{1}{Z} e^{\beta \sum_{i=1}^N M(v_i, v_{i+1})} \quad (14)$$

where the interaction kernel M mimics the XY energy function,

$$M(v, v') = \cos\left(\frac{2\pi}{P}(v - v')\right), \quad (15)$$

and the partition function normalizes the distribution p . This distribution enjoys two symmetries, compare to the single symmetry of the Ising model in (10): for any integers K and L we

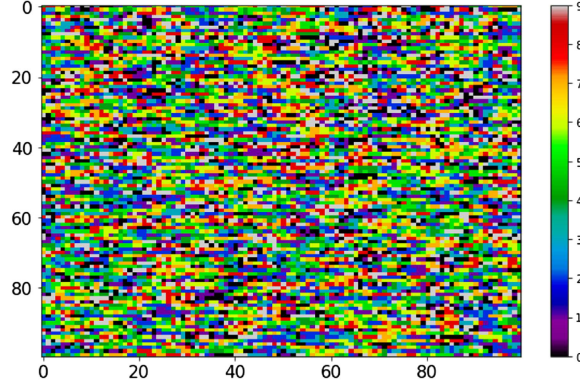


Figure 8. 100 configurations (index along the y-axis) of the discretized XY model with $P = 10$ Potts states over $N = 100$ visible units (x-axis, index i of the units) at inverse temperature $\beta = 1.5$. Each color refers to one of the 10 Potts states $v = 0, 1, \dots, 9$.

have,

$$p_{\text{data}}(v_1, v_2, \dots, v_N) = p_{\text{data}}(v_{k+1} + L, v_{k+2} + L, \dots, v_{k+N} + L), \quad (16)$$

where $i + k$ and $v + L$ are to be intended, respectively, modulo N and P . Figure 8 shows a set of 100 configurations over $N = 100$ sites, generated independently and at random from this model for $P = 10$.

4.2. Symmetry-breaking in both spaces

4.2.1. Case of a single hidden unit. We consider a RBM with $N = 100$ visible Potts-type units v_i , which can take one out of $P = 10$ values, and with $M = 1$ hidden unit. The weights $w_{i,\mu=1}$ is now a vector $w_{i,1,v}$, with $i \in [1, 2, \dots, 100]$ and $v \in [1, 2, \dots, 10]$. The component $w_{i,1,v}$ of this vector is the connection between the hidden unit and the visible unit i when it carries the Potts state v .

We first train a RBM with a single hidden unit h_1 , which takes real values and is submitted to a double-well potential. Figure 9(a) shows the weights obtained after training from a very large number of configurations, starting from small white noise initial conditions for the $w_{i,1,v}$. We observe a strong modulation of the weights in the space and angle directions, achieving peak values around some site i and angle v . Similar results were found for a binary-value hidden unit, $h_1 = \pm 1$, with a slightly weaker localization of the weights and at a different location, see figure 9(b). In the following, we show results obtained for the RBM with the real-value hidden unit only.

Since the interaction matrix M in the Potts model takes the cosine function form, our RBM should learn the same functional dependence from the data samples. We show in figure 9(c) the quantity

$$W_i^{\text{angular}} = \left| \sum_{v=0}^{P-1} w_{i,1,v} \right|, \quad (17)$$

which measures the angular modulation of the weights on each site i . We see a strong space localization around $i = 39$, because the weights only take non-zero values near that location. This location is arbitrary and similar to the place-field formation accompanying the breaking

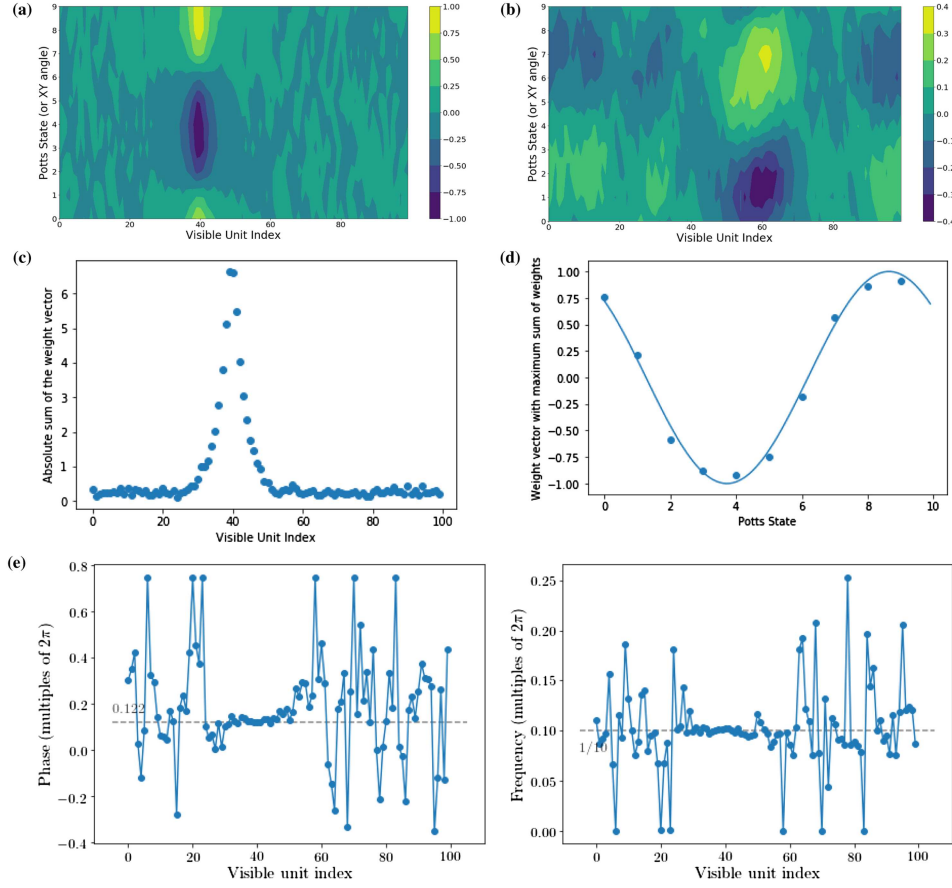


Figure 9. (a) Contour plot of the weights of the RBM with a single real-valued hidden unit with double-well potential trained on the XY model discretized by the Potts model. The x -axis shows space (index i of the visible units), while the y -axis refers to angles (Potts state v). Parameters: $P = 10$ Potts states, $N = 100$ visible units, $M = 1$ hidden unit, trained on 100 000 configurations, learning rate $\nu = 0.01$ and batch size $S = 100$ trained over 100 epochs. (b) Same as (a) for a binary hidden unit. (c) Angular modulation W_i^{angular} of the weight vector as a function of the space location i , see (17). Same parameters as in figure 9(b) for the RBM with real-valued hidden unit. There is a clear strong space localization with a peak centered in unit $i = 39$. (d) Weight vector $w_{i=39,\mu=1,v}$ as a function of the angular-Potts variable v . The line represents the cosine function with frequency $= 2\pi/10$ as expected, with the best fit of the phase. (e) Phases φ_i vs site index i . Gray dotted line is the phase of the above fitting cosine. (Right) Frequency ω_i vs site index i . Gray dotted line is again the frequency of $2\pi/10$ of the cosine fit above, which is what one would expect from system with 10 Potts states. See text for the definition of the fitted frequencies and phases. The phase and the frequency is constant across the size of the receptive field, that is all the spins look in the same direction.

of translation symmetry over space observed for the Ising model. In addition, at the location of the maxima, the weight vector is very well approximated by a cosine function, see figure 9(d). The RBM has learned the correct frequency equal to $2\pi/10$, and the phase takes an arbitrary value. Indeed, the phase is a free parameter due to the invariance against choices of L in (16).

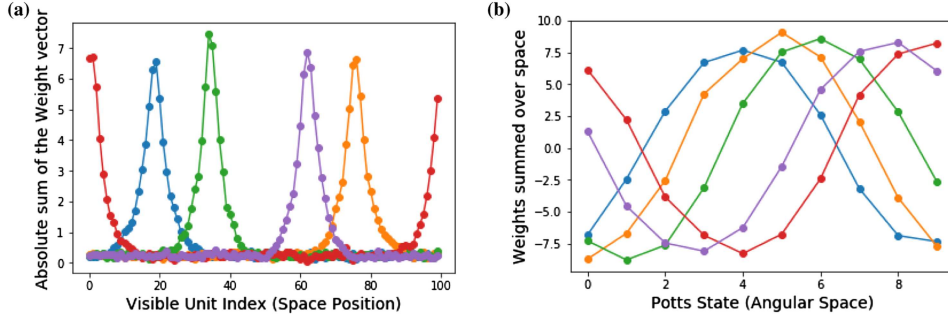


Figure 10. (a) Angular modulation $W_{\mu,i}^{\text{angular}}$ (similar to (17), but the hidden-unit index μ runs from 1 to 5) vs space location i . All the parameters are same as those of figure 9 but with $M = 5$ real valued hidden unit with double well potential. There is a clear strong space localization and the receptive fields for the different units show mutual repulsion of their weight vectors. (b) Weight vectors at the maximum visible unit index $i_{\max}(\mu)$ for the respective hidden units μ . The different curves each lie on a cosine function with frequency $= 2\pi/10$ as expected but with phases showing equal separations when ranked in increasing order.

To obtain a more precise picture of the receptive field, we then consider, for each site i , the P -dimensional vector of the weights $w_{i,\mu=1,v}$. We then fit this vector with a cosine function of adjustable frequency and phase, referred to as, respectively, ω_i and φ_i . We show, as functions of the site index i , the periods ω_i and the phases φ_i in figure 9(e). We observe that the period takes the expected value $2\pi/P$ over the receptive field (sites ranging approximately between $i = 30$ and 50). Similarly, the phase is constant (and takes an arbitrary value) over the same region of space. Informally speaking, when the hidden unit is on, all the XY spins supported by the sites in the receptive field point to the same direction.

4.2.2. Case of multiple hidden units. We also train a RBM with $M = 5$ real valued hidden unit with double well potential, with results shown in figure 10. We see that the receptive fields of the hidden units are mutually separated in space, and show the same phenomenon of repulsion between the units observed for the Ising data. In addition, the angular dependence of the five weight vectors exhibit the same frequency (equal to $2\pi/10$), but the phases show also a nice equi-separation due to repulsion along the angular direction.

Though we expected to see a diffusion of the receptive fields both along the spatial and angular dimensions for very large learning times, we did not observe this phenomenon even with RBM trained with 1000 000 samples. This is likely due to the fact that the landscape is still rough for this amount of data, and diffusion remains activated. We have not tried to increase the number of samples because of the high computational cost.

4.3. Differentiated retarded learning transitions

In this section, we show that RBM trained with data generated by the discretized XY model shows retarded learning phase transitions. However, as there are two potential symmetry breaking directions in this model, one corresponding to the angular space and the other to the positional space, the breaking of symmetry along these direction may take place at two different values of the noise ratio $r = N/S$, i.e. for different number of samples in the data set used for training. The reason is that the number of Potts states in the angular direction, P , may largely differ from the number of sites on the lattice, N . Consequently, the effective system sizes along the two directions are different.

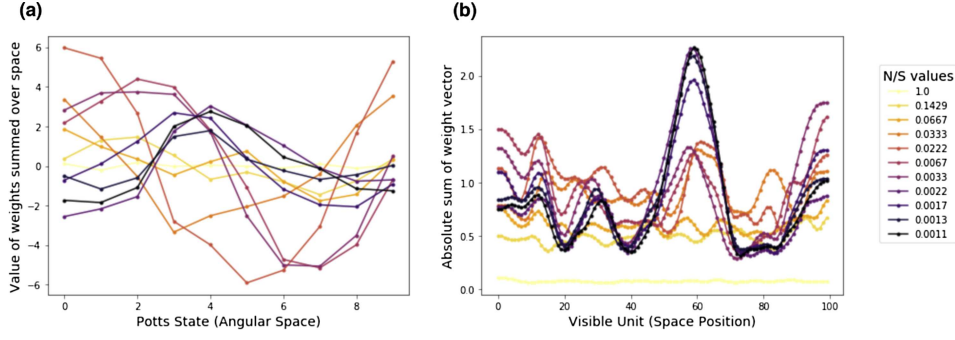


Figure 11. The phenomenon of differentiated retarded learning phase transitions, for various noise ratios $r = N/S$, where S is the number of training samples. (a) Angular direction: The y-axis is the sum of the components $w_{1,i,v}$ of the weight vector over all spatial locations $i = 1, \dots, N$. The x-axis shows the discretized angular states $v = 0, 1, \dots, P - 1$. (b) Spatial direction: The y-axis is the sum of the components $w_{1,i,v}$ of the weight vector over all angular variables $v = 1, \dots, P$. The x-axis shows the lattice site $i = 1, 2, \dots, N$. Parameters for data generation and sampling: $N = 100$, $P = 10$, $\beta = 1.5$, $M = 1$ (Bernoulli hidden unit), $\nu = 0.1$, $S_{\text{batch}} = 100$, trained for 20 epochs.

This phenomenon of differentiated retarded learning phase transitions is reported in figure 11. We show in panel (a) of the figure the spatial modulation defined through,

$$W_v^{\text{spatial}} = \sum_{i=1}^N w_{i,1,v}, \quad (18)$$

as a function of the Potts angular state variable v . We observe that for a large r , the spatial modulation vanishes all over the angular space: low amount of data are not sufficient for the RBM to capture the angular correlations in the configurations. For large enough data set ($r < 0.033$) the spatial modulation shows a clear dependence on v . We then show in panel (b) of figure 11 the angular modulation W_i^{angular} as a function of the lattice site index i for varied levels of sampling noise, r . Again, for large r , no modulation is seen. However, for very small noise levels $r < 0.002$, we do observe that W_i^{angular} is peaked around some well defined site i . Interestingly, in the range $0.002 < r < 0.008$, the angular modulation does not significantly vary over space, while the spatial modulation varies over angles, compare panels (a) and (b). We conclude that, for intermediate ranges of values of r , the RBM has created a place-field along the angular direction, but not along the spatial direction.

To test the generality of the phenomenon of differentiated transitions, we also generated data samples from variants of the discretized XY model. We modified the XY model in terms of changing the interaction matrix M in (14) from the cosine function to short range couplings, and also we changed the Hamiltonian to include not only nearest neighbor couplings but also long range couplings in the positional space. The resulting models display a variety of phase transitions in the RBM weights after training, with positional symmetry breaking arising before (for smaller amount of training data) angular ordering in some cases (not shown).

5. Theoretical analysis

Hereafter, we study analytically the dynamics of learning of the weights of the RBM with binary hidden units when trained with data. Two limit cases will be considered:

- The case of few data, which allows us to establish the connection with random matrix theory and the so-called retarded learning transition;
- The case of a large amount of data, with weak correlations, which we analyze in detail to understand the formation and shape of the place field, as well as the interactions between different place fields arising through learning.

While we will focus on the learning dynamics of the weights, we assume that the RBM has correctly learned the local fields, so we will set $b_i = c_\mu = 0$ from the beginning in the case of unbiased binary data $v_i = \pm 1$. In addition, we assume that hidden units are of Bernoulli type, $h_\mu = \pm 1$. The log-likelihood therefore reads

$$\log \mathcal{L} = \left\langle \sum_{\mu=1}^M \log \cosh \left(\sum_{i=1}^N w_{i\mu} v_i \right) \right\rangle_{\text{data}} - \log \mathcal{Z}(\{w_{i\mu}\}), \quad (19)$$

where the partition function is

$$\mathcal{Z}(\{w_{i\mu}\}) = \sum_{\{v_1, v_2, \dots, v_N\}} \prod_{\mu=1}^M \cosh \left(\sum_{i=1}^N w_{i\mu} v_i \right). \quad (20)$$

Taking the partial derivative with respect to $w_{\mu i}$ we get the following expression for the gradient of the log-likelihood:

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial w_{i\mu}} &= \left\langle v_i \tanh \left(\sum_{j=1}^N w_{j\mu} v_j \right) \right\rangle_{\text{data}} - \frac{1}{\mathcal{Z}(\{w_{i\mu}\})} \sum_{\{v_1, v_2, \dots, v_N\}} v_i \\ &\quad \times \sinh \left(\sum_{j=1}^N w_{j\mu} v_j \right) \prod_{\lambda(\neq \mu)} \cosh \left(\sum_{j=1}^N w_{i\lambda} v_j \right). \end{aligned} \quad (21)$$

The continuous-time dynamical equations for the evolution of the weights during training, assuming that the batch size is maximal, i.e. that all the data are used for training, are

$$\frac{dw_{i\mu}}{dt} = \nu \frac{\partial \log \mathcal{L}}{\partial w_{i\mu}}, \quad (22)$$

where ν is the learning rate.

5.1. Few data: small weight expansion and the retarded learning transition

5.1.1. Linearized equations of the dynamics. In this section, we assume that the weights have initially very small (random) values. For small enough learning times, we may linearize the dynamical equation (22). We obtain

$$\frac{dw_{i\mu}}{dt} = \nu \left(\sum_{j=1}^N C_{ij} w_{j\mu} - w_{i\mu} \right), \quad (23)$$

where

$$C_{ij} = \langle v_i v_j \rangle_{\text{data}} \quad (24)$$

is the empirical covariance matrix estimated from the data. Let Λ be the largest eigenvalue of C , and \mathbf{e} the associated eigenvector, with components e_i . As the diagonal elements C_{ii} are equal to unity ($v_i^2 = 1$), we have that $\Lambda > 1$, unless C is the identity matrix and the data shows no correlation at all. Hence, according to (23), all weight vectors $\mathbf{w}_\mu = \{w_{1\mu}, w_{2\mu}, \dots, w_{N\mu}\}$ align along \mathbf{e} ; this result holds within the linear approximation, and is therefore expected to be valid at short times only.

Let us consider the noise ratio $r = N/S$, equal to the number of visible units (system size) over the number of training samples. For bad sampling (large r), the empirical covariance matrix can be approximated by the covariance matrix of a null model, in which all N visible units are independent and unbiased: v_i is equal to ± 1 with equal probabilities ($= 1/2$), independently of the other v_j 's. The asymptotic distribution of the eigenvalues of such a random matrix has a special form, called the Marcenko Pastur (MP) spectrum [30], whose right edge (top eigenvalue) is given by

$$\Lambda_{\text{noise}} = \Lambda_{\text{MP}} = (1 + \sqrt{r})^2 \quad (25)$$

and the corresponding top eigenvector \mathbf{e} has random, Gaussian distributed components.

Conversely, for good sampling (small r), we expect the empirical covariance to be similar to the covariance matrix computed from the model distribution p from which data were generated. Due to the translational invariance of p , its top eigenvector $\mathbf{e}_{\text{model}}$ has the same symmetry: $\mathbf{e}_{\text{model}} = (1, 1, \dots, 1)$, up to a normalization factor. Hence, we expect \mathbf{e} to be similar to $\mathbf{e}_{\text{model}}$ and be roughly uniform. In the double, large N and large S limit, the two regimes may be separated by a sharp transition, taking place at a critical value of r . To locate this value, we compute below the top eigenvalue of the model covariance matrix, and compare it to its MP counterpart (25). The crossover between the bad and good sampling regimes takes place when both eigenvalues are equal.

5.1.2. Case of Ising data. Let us consider the case of the one-dimensional Ising model. When a large number of configurations is available, we have $C_{ij} = (\tanh \beta)^{|i-j|}$, see (9). Due to the rotational invariance, the top-eigenvector, $\mathbf{e}_{\text{Ising}}$ has all its components equal. Therefore the top eigenvalue of the covariance matrix is

$$\Lambda_{\text{Ising}}(\beta) = \sum_{j=1}^N C_{ij} = 1 + 2 [\tanh \beta + \tanh^2 \beta + \tanh^3 \beta + \dots] \approx \frac{1 + \tanh \beta}{1 - \tanh \beta} = e^{2\beta}. \quad (26)$$

When the inverse temperature β is small, this ‘signal’ eigenvalue is smaller than the ‘noise’ eigenvalue Λ_{MP} (25), locating the right edge of the MP spectrum. In this case, we expect the top eigenvector \mathbf{e} of the empirical covariance matrix C to be noisy, and not to capture the correlation between the Ising variables v_i . In this regime, no receptive field with a localized weight structure can emerge. As β increases above

$$\beta(r) = \log(1 + \sqrt{r}), \quad (27)$$

the signal eigenvalue $\Lambda_{\text{Ising}}(\beta)$ becomes larger than the MP edge, and we expect the top eigenvector of C to have comparable component and be similar to $\mathbf{e}_{\text{Ising}}$.

The above statement is corroborated by the results shown in figure 12 (a,b), which shows the top eigenvalue Λ of the correlation matrix C (24) as a function of the noise ratio, $r = N/S$, where S is the number of samples. For large r (few samples), Λ is very well approximated

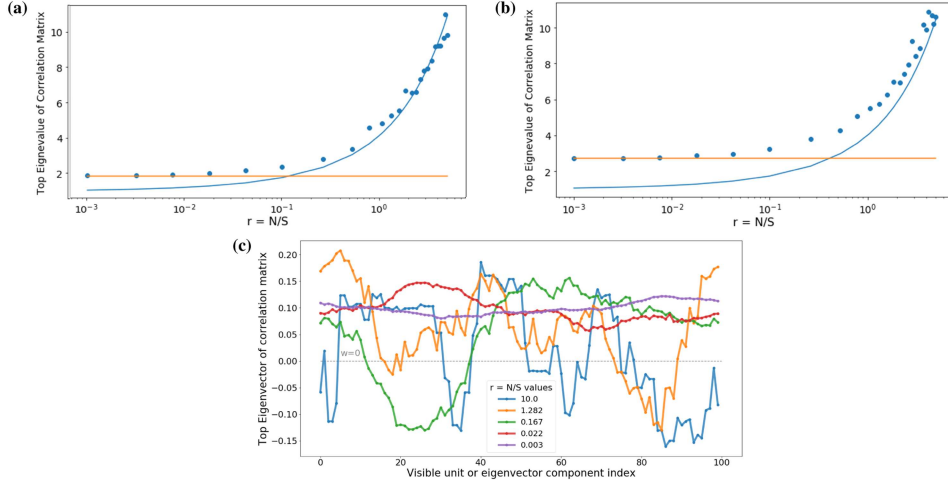


Figure 12. (a) Top Eigenvalue of the correlation matrix C of data generated from the one-dimensional Ising model at inverse temperature $\beta = 0.3$, as a function of the noise ratio r . The orange straight line is the top eigenvalue Λ_{Ising} , corresponding to a perfectly sampled (infinite S) Ising model. The blue curve is the top eigenvalue Λ_{noise} of the correlation matrix of the null model with independent variables. (b) Same as panel (a) but for $\beta = 0.5$. (c) Top eigenvector of the correlation matrix C (24) of the configurations of the one-dimensional Ising model at fixed β but with different numbers S of samples. Ising model samples to calculate the correlation matrix were generated at $\beta = 1$ for $N = 100$ spins.

by Λ_{noise} , while, for small r (many samples), Λ gets very close to $\Lambda_{\text{Ising}}(\beta)$ as expected. The crossover between these two regimes takes place at values of r such that $\beta \simeq \beta(r)$ (27).

Figure 12(c) shows how the top eigenvector of the data correlation matrix changes as more and more samples are considered. One clearly sees a phase transition from a random vector to the uniform eigenvector $\mathbf{e}_{\text{Ising}}$.

5.1.3. Case of XY data. For the discrete XY model, the correlation matrix in the $r \rightarrow 0$ limit can be computed as well using the transfer matrix formalism. We find

$$C_{ij}(v, v') = \frac{1}{P-1} \sum_{p=1}^{P-1} \left(\frac{\lambda_p(\beta)}{\lambda_0(\beta)} \right)^{|j-i|} \cos \left(\frac{2\pi p(v-v')}{P} \right), \quad (28)$$

where

$$\lambda_p(\beta) = \sum_{v=0}^{P-1} \exp \left[\beta \cos \left(\frac{2\pi v}{P} \right) \right] \cos \left(\frac{2\pi v p}{P} \right). \quad (29)$$

C enjoys translational invariance along both axis, hence its eigenvectors are discrete 2D Fourier modes; after computation, we find that the top eigenvalue is

$$\Lambda_{\text{XY}}(\beta) = \frac{P}{P-1} \frac{\lambda_0(\beta) + \lambda_1(\beta)}{\lambda_0(\beta) - \lambda_1(\beta)}. \quad (30)$$

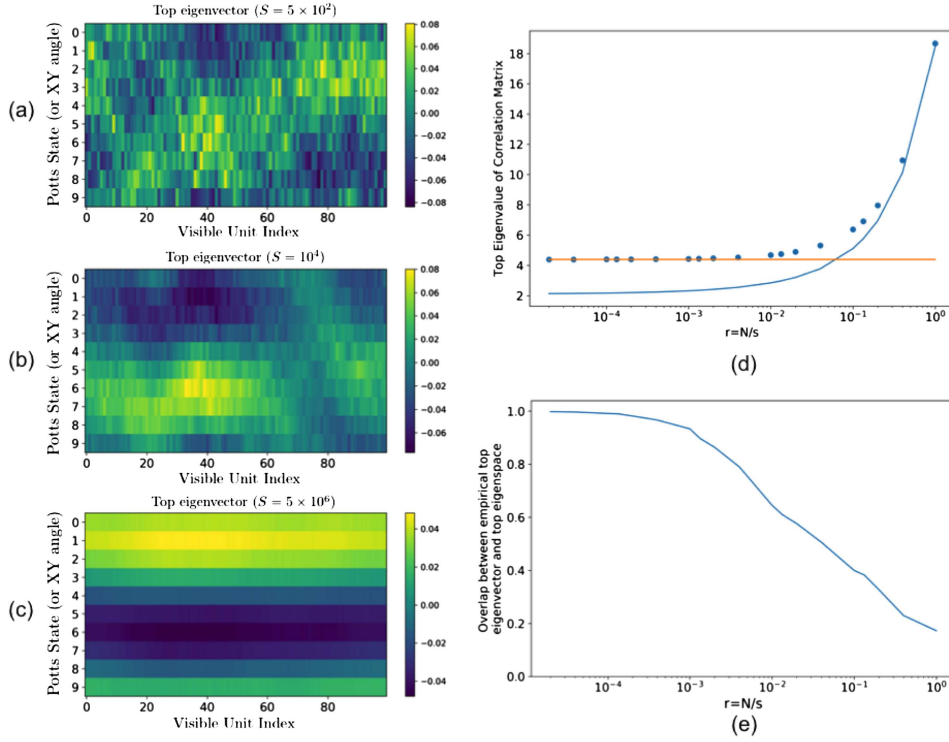


Figure 13. (a)–(c) Empirical top eigenvector of the correlation matrix C of data generated from the XY model at inverse temperature $\beta = 1.5$, for sample size $S = 5 \times 10^2$ (a), $S = 10^4$ (b), $S = 5 \times 10^6$ (c). (d) Corresponding top eigenvalue (dots) as a function of the sample size $r = \frac{N}{S}$. The orange straight line is the top eigenvalue $\Lambda_{XY}(\beta = 1.5)$, corresponding to a perfectly sampled (infinite S) XY model. The blue curve is the top eigenvalue Λ_{noise} of the correlation matrix of the null model with independent variables. The eigenvalues cross $r \simeq 0.066$. (e) Norm of the projection of the empirical top eigenvector $\hat{\mathbf{e}}$ in the space spanned by the perfect-sampling top eigenvectors $\mathbf{e}^1, \mathbf{e}^2$, $\sqrt{(\hat{\mathbf{e}} \cdot \mathbf{e}^1)^2 + (\hat{\mathbf{e}} \cdot \mathbf{e}^2)^2}$.

with a corresponding eigenspace of dimension 2, spanned by $\mathbf{e}_i^1(v) = \sqrt{\frac{2}{NP}} \cos\left(\frac{2\pi v}{P}\right)$, $\mathbf{e}_i^2(v) = \sqrt{\frac{2}{NP}} \sin\left(\frac{2\pi v}{P}\right)$. The top eigenvector is uniform over space, as for the Ising model, but not over the angular variables, see figure 13(c).

The ‘noise’ eigenvalue is similarly given by the MP spectrum, although slightly modified: the dimension to sample size ratio is now $\frac{PN}{S} = Pr$, and in the $S \rightarrow \infty$ limit, the correlation matrix has top eigenvalue $\frac{P}{P-1}$ owing to the anticorrelations between Potts variables on the same site, $C_{i,i}(a,b) = -\frac{1}{P-1}$, $\forall a \neq b$. We obtain:

$$\Lambda_{\text{noise}} = \frac{P}{P-1} + \left(1 + \sqrt{rP}\right)^2 \quad (31)$$

Similarly to the case of Ising data, when β is small, the signal ‘eigenvalue’ is small compared to the ‘noise’ eigenvalue, and the empirical top eigenvector has a small projection in the space spanned by $\mathbf{e}^1, \mathbf{e}^2$, see figure 13(b)–(d). The crossover between the two regimes takes place at

values of r such that $\Lambda_{MP} \simeq \Lambda_{XY}(\beta)$. The first retarded learning transition of the RBM occurs in the same range of r , see figure 11.

5.2. Many data: small β expansion

After some training time, linear equation (23) for the weights breaks down and non linearities must be taken into account [31]. We derive below an approximation to the RBM dynamic learning equation (with $M = 1$ or 2 hidden units) for the one-dimensional Ising models, which is exact for small (but non vanishing) inverse temperature β . We show that this equation is free of any external parameters after appropriate rescaling of the weights. We compare the numerical solutions to this equation with the result of the training with RBM to find a parameter independent agreement with the shape and the structure of the weights. We also cast the equation into a continuous form, and formulate the system in terms of a standard reaction–diffusion instability problem with the weights as an inducer and the sum of weights squared as the repressor.

5.2.1. One hidden unit system: formation of receptive field. For one hidden unit, equations (21) and (22) become, after some elementary manipulation,

$$\frac{\partial \log \mathcal{L}}{\partial w_j} = \left\langle v_i \tanh \left(\sum_{j=1}^N w_j v_j \right) \right\rangle_{\text{data}} - \tanh w_i, \quad (32)$$

where we have dropped the $\mu = 1$ index for the sake of clarity. Expanding the hyperbolic tangents to the third powers of their arguments, we obtain

$$\frac{\partial \log \mathcal{L}}{\partial w_j} = \sum_j \langle v_i v_j \rangle_{\text{data}} w_j - \frac{1}{3} \sum_{j,k,l} \langle v_i v_j v_k v_l \rangle_{\text{data}} w_j w_k w_l - w_i + \frac{1}{3} w_i^3 + O(w^4). \quad (33)$$

Let us now assume that a large number of samples is available. At the lowest order in β , we have

$$\langle v_i v_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ \beta & \text{if } i = j \pm 1, \\ 0 & \text{otherwise.} \end{cases} \quad (34)$$

and

$$\langle v_i v_j v_k v_l \rangle = \begin{cases} 1 & \text{if } i = j, k = l \text{ or any permutation,} \\ \beta & \text{if } i = j \pm 1, k = l \text{ or any permutation,} \\ \beta & \text{if } i = j, k = l \pm 1 \text{ or any permutation,} \\ 0 & \text{otherwise.} \end{cases} \quad (35)$$

for, respectively, the 2- and 4-point correlations. We therefore obtain

$$\frac{\partial \log \mathcal{L}}{\partial w_i} = \beta(w_{i+1} + w_{i-1}) - w_i \sum_k w_k^2 + w_i^3 + O(w^4, \beta w^3). \quad (36)$$

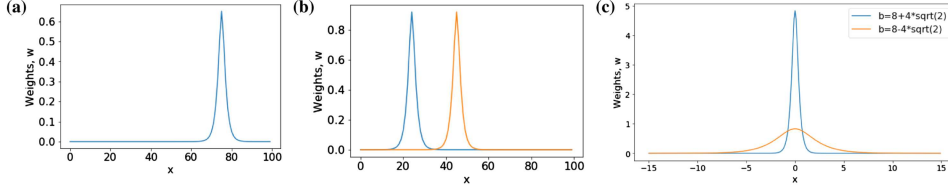


Figure 14. (a) Stationary solution of the small β equation (37), describing the evolution of the weights $w_i(t)$ of a RBM with $M = 1$ hidden unit trained over many configurations of the one-dimensional Ising model. There are $N = 100$ visible units. The results shown were obtained with 500 integration steps, starting from small amplitude white noise initial conditions for the weights. (b) Same as (a), but for a RBM with $M = 2$ hidden units. The two colors shows the weights corresponding to the two units. While the two peaks should be in principle diametrically opposed, i.e. at distance 50 from each other, their mutual repulsion is short ranged; in practice deviations from stationarity smaller than the numerical accuracy cannot be detected. (c) Profiles of the stationary weight vector for a RBM with a unique hidden unit trained on data extracted from the one-dimensional Ising model at small inverse temperature β , see text. The two curves corresponds to the two candidate values for b . The solutions $b = b_+$ and $b = b_-$ are, respectively, unstable and stable against small fluctuations of the weights.

Upon appropriate rescaling of the weights, $w_i \rightarrow w_i/\sqrt{\beta}$, and of the learning rate, $\nu \rightarrow \nu/\beta$, we obtain, in the small β regime, the non trivial, parameter-free dynamical equation

$$\frac{1}{\nu} \frac{dw_i}{dt} = w_{i+1} + w_{i-1} - w_i \sum_k w_k^2 + w_i^3. \quad (37)$$

The stationary solution of this equation is shown in figure 14(a).

This equation can be cast in a continuous form over space, where we use the Laplacian to describe spatial diffusion. The corresponding continuous partial differential equation reads

$$\frac{1}{\nu} \frac{\partial w}{\partial t}(x, t) = \frac{\partial^2 w}{\partial x^2}(x, t) + (2 - b(t))w(x, t) + w(x, t)^3, \quad (38)$$

where

$$b(t) = \int_0^L w(x, t)^2 dx. \quad (39)$$

These coupled dynamical equations lead to a non-trivial spatial formation through the so-called *Turing reaction diffusion instability* mechanism [32]. The field $w(x, t)$ diffuses over space and activates itself (self-promoting, through the cubic term), but is inhibited by another species, b . This repressor is diffusing with an infinite diffusion coefficient, i.e. is spatially uniform, and depends on w through (39). As w grows due to self-activation, so does the repressor b , until w reaches a stationary profile. We show in appendix that the above dynamical equation satisfy the general criteria for stable pattern formation.

5.2.2. One hidden unit: profile of the receptive field. Consider the stationary continuous equation satisfied by the weights after learning, see (38),

$$0 = \frac{d^2 w}{dx^2}(x) + (2 - b) w(x) + w(x)^3. \quad (40)$$

Multiplying by $\frac{dw}{dx}$ on both sides and integrating over x we obtain that

$$E(x) \equiv \frac{1}{2} \left(\frac{dw}{dx} \right)^2 + w(x)^2 + \frac{1}{4} w(x)^4 - \frac{b}{2} w(x)^2 \quad (41)$$

has a uniform value E_0 , independent of x . When $x \rightarrow \pm\infty$, both $w(x)$ and $\frac{dw}{dx}(x)$ tend to 0, which sets $E_0 = 0$. We deduce that

$$\frac{dw}{dx}(x) = \pm w(x) \sqrt{b - 2 - \frac{w(x)^2}{2}}. \quad (42)$$

We now explicitly break the symmetry by fixing the center of the peak of the weights in $x = 0$, with $w(0) > 0$, $\frac{dw}{dx} > 0$ for $x < 0$, and $\frac{dw}{dx} < 0$ for $x > 0$. Imposing that the derivative of the weight with respect to x vanishes at its maximum, i.e. that w is twice differentiable in $x = 0$ gives

$$w(0) = \sqrt{2(b - 2)}. \quad (43)$$

Integrating (42) with condition (43), we find

$$w(x) = \frac{\sqrt{2(b - 2)}}{\cosh(x\sqrt{b - 2})}. \quad (44)$$

Using definition (39) for b we then find

$$b = \int_{-\infty}^{\infty} w(x)^2 dx = 4\sqrt{b - 2}, \quad (45)$$

whose solutions are $b_{\pm} = 8 \pm 4\sqrt{2}$. The corresponding profiles of the weights are shown in figure 14(c). We now study the stability of the solution under the time-dependent perturbation $w(x) \rightarrow w(x) + \epsilon(x, t)$, where $w(x)$ is given by (44). According to equations (38) and (40), we have

$$\frac{1}{\nu} \frac{\partial \epsilon}{\partial t}(x, t) = \frac{\partial^2 \epsilon}{\partial x^2}(x, t) + (2 - b) \epsilon(x, t) - 2 \left(\int dy w(y) \epsilon(y, t) \right) \epsilon(x, t) + 3 w(x)^2 \epsilon(x, t). \quad (46)$$

Multiplying by $w(x)$ and integrating over x , we get the following equation

$$\frac{1}{\nu} \frac{d}{dt} \int dx w(x) \epsilon(x, t) = -\sqrt{b - 2} \int dx w(x) \epsilon(x, t) \left(8 - \frac{b}{\cosh(x\sqrt{b - 2})^2} \right). \quad (47)$$

We deduce that the weight profile is stable if and only if $b < 8$. Therefore, the $b = b_+$ solution is unstable against small variations of the peak amplitude near $x = 0$, and the solution $b = b_-$ is the correct, stable one. Notice that the width of the peak of the weight, in the $\beta \rightarrow 0$ limit is finite according to expression (44). This phenomenon was also observed by the RBM training results in figure 3(b), where the peak width obtained by linear fit (coefficient b) was also positive and finite.

5.2.3. Two hidden units: interaction between receptive fields. For two hidden units, equations (21) and (22) become, after some simple manipulation,

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial w_{j1}} = & \left\langle v_i \tanh \left(\sum_{j=1}^N w_{j1} v_j \right) \right\rangle_{\text{data}} - \frac{\tanh(w_{i1} + w_{i2})}{1 + \prod_{k=1}^N \frac{\cosh(w_{k1} - w_{k2})}{\cosh(w_{k2} + w_{k2})}} \\ & - \frac{\tanh(w_{i1} - w_{i2})}{1 + \prod_{k=1}^N \frac{\cosh(w_{k1} + w_{k2})}{\cosh(w_{k1} - w_{k2})}}, \end{aligned} \quad (48)$$

together with a similar equation for the weight vector $\mu = 2$ obtained by swapping the hidden-unit indices 1 and 2. Notice that this equation simplifies to (32) when the weight vector $\mu = 2$ is set to zero, i.e. when $w_{i2} = 0$ for all visible units i , and the number of hidden units is effectively $M = 1$.

Let us now expand (48) in powers of the weights. The first term on the right hand side of the equation (involving the average over the data distribution) has the same expansion as in the $M = 1$ case above, see (33). For the second term, using

$$1 + \prod_{k=1}^N \frac{\cosh(w_{k1} - w_{k2})}{\cosh(w_{k1} + w_{k2})} = 2 - 2 \sum_{k=1}^N w_{k1} w_{k2} + O(w^4), \quad (49)$$

and rescaling the weights, $w \rightarrow w/\sqrt{\beta}$, and the learning rate, $\nu \rightarrow \nu/\beta$, as before, we obtain

$$\frac{1}{\nu} \frac{dw_{i1}}{dt} = w_{i+1,1} + w_{i-1,1} - w_{i1} \sum_k w_{k1}^2 + w_{i1}^3 + w_{i1} w_{i2}^2 - w_{i2} \sum_k w_{k1} w_{k2}. \quad (50)$$

Similarly, we find

$$\frac{1}{\nu} \frac{dw_{i2}}{dt} = w_{i+1,2} + w_{i-1,2} - w_{i2} \sum_k w_{k2}^2 + w_{i2}^3 + w_{i2} w_{i1}^2 - w_{i1} \sum_k w_{k1} w_{k2}. \quad (51)$$

The last two terms in the two equations above encode the couplings between the weight vectors attached to the two units. The stationary solutions of these equations are shown in figure 14(b). In practice, we observe that the numerical solutions for the weight profiles attached to the two units can have any relative separation between their peaks as long as it is larger than approximately one peak width. The repulsion between the peaks is indeed short range, hence the convergence to the diametrically opposed configuration is really slow.

These equations can be turned into two partial differential equations over continuous space where we resort to the Laplacian to describe spatial diffusion:

$$\begin{aligned} \frac{1}{\nu} \frac{\partial w_1}{\partial t}(x, t) &= \frac{\partial^2 w_1}{\partial x^2}(x, t) + (2 - b_1(t)) w_1(x, t) - c w_2(x, t) + w_1(x, t) (w_1(x, t)^2 + w_2(x, t)^2), \\ \frac{1}{\nu} \frac{\partial w_2}{\partial t}(x, t) &= \frac{\partial^2 w_2}{\partial x^2}(x, t) + (2 - b_2(t)) w_2(x, t) - c w_1(x, t) + w_2(x, t) (w_1(x, t)^2 + w_2(x, t)^2), \end{aligned} \quad (52)$$

where

$$b_1(t) = \int_0^L w_1(x, t)^2 dx, \quad b_2(t) = \int_0^L w_2(x, t)^2 dx, \quad c(t) = \int_0^L w_1(x, t) w_2(x, t) dx, \quad (53)$$

This system describes two diffusing species w_1 and w_2 which are, respectively, self-inhibited by b_1 and b_2 , while coefficient c corresponds to cross-inhibition. The diffusion coefficient for b_1, b_2, c tend to infinity and their concentrations are uniform in space. As in the single species (single hidden-unit) case, this dynamical system leads to the stable production of a non trivial pattern over space, corresponding to the emergence of two place fields, see [appendix](#).

6. Conclusion

In this work, we have studied the unsupervised learning of simple data distributions, enjoying one or two continuous symmetries, with a RBM. Contrary to standard approaches in machine learning, e.g. convolutional networks, we have not tried to factor out, and hardwire these symmetries in the network. On the contrary, our objective was to see how the symmetries affected the representations of the data and were effectively learned by the machine. This approach is motivated by the fact that most invariances in complex data are actually unknown and it is important to understand how well they can be captured in practice.

In the case of a single hidden (latent) variable, our main observation is that learning is accompanied by a symmetry breaking in the weight space. The hidden unit concentrates only on a small portion of the data configurations; the size of this receptive field is the length over which the variables in the data configurations are correlated. The symmetry is dynamically restored at long times through the diffusion of the receptive field, allowing it to span the whole data manifold⁵. This phenomenon is strongly reminiscent of the concept of continuous attractor (CA) in the context of recurrent neural networks in computational neuroscience [33–35], with the major differences that (1) CA usually refer to low-dimensional attractors in the (high-dimensional) space of neural activities, while the CA emerging here defines a low-dimensional manifold in the *weight space*, and (2) accordingly, the dynamics considered is the *learning* dynamics acting on weights and not the usual neural dynamics modifying activities. In the case of multiple hidden units, the CAs attached to these units are locked in: weight bumps diffuse coherently along their CA's (figure 6), maintaining their relative phases due to mutual repulsive interactions. The resulting multi-unit CA has therefore the same (low) dimension as the underlying symmetry in the data. In practice, however, repulsion is short ranged and may effectively lead to partial decoupling, see figure 14(b), and to an increase in the CA dimension. If the number of hidden units is sufficiently large (of the order of the number of visible units over the correlation length) the RBM hidden configurations are effective, coarsegrained version of the data configurations.

An important condition for this CA in the weight space to emerge is that the number of available data exceeds some critical value depending on the configuration size and the intensity of their intrinsic correlations. This phenomenon is a manifestation of the general mechanism of the so-called retarded-learning phase transitions [16], in which a symmetry-breaking direction

⁵The fact that the bump or the peak of weights can be localized anywhere in the visible space is a direct consequence of the translation invariance of the underlying data; we have checked that, with non-translationally invariant couplings, the bump is pinned by the strongest interactions.

(here, in the weight space) is inferred when the ratio of the number of data and of the system size is larger than some signal-dependent (here, the level of intrinsic correlation in the data) threshold. Interestingly, in the presence of multiple invariances, the thresholds associated to these symmetries need not coincide. In such situations, the receptive field will be localized along one dimension in the input space and extended along the other, as seen for the XY model in this work.

It is tempting to make an analogy with recent experimental results on three-dimensional encoding by place cells [36]. When a rodent explores a set of horizontal (x direction) and vertical (z direction) planes, place cells emerge with place fields localized in either or both planes [37]. Yet, if the motion along the x and z axis is not independent, localization can be lost along one of the two directions. For instance, when motion takes place along a helicoidal ramp (x being the angle in the plane perpendicular to the helix axis z), place cells seem to be localized in the angular space and much less so along the vertical axis [38]. Due to the geometry of the helix, it is reasonable to assume that inputs related to path integration as well as to visual flow are strongly correlated for similar angles (corresponding to a small displacement on the ramp) and much less correlated for small translation along the z -axis, which requires a large physical displacement. It would be interesting to see what happens if the ramp axis is tilted and not vertical any longer. Based on the analogy with the differential retarded-learning transitions, one would expect that place fields are columnar along the ramp axis, and become therefore localized (albeit with different areas) along both x and z directions.

While the analogy with place cells and symmetry-broken hidden units is tempting, establishing a solid connection between our results and neuroscience is far from obvious. Though place cells are known to rely, for their establishment, on various sources of sensory information (including visual and path-integration inputs [39]), the mechanisms underlying the corresponding unsupervised learning processes are far from being elucidated. It is, from this point of view, remarkable that various unsupervised learning rules [6, 11, 12] agree with the two main features emerging from the maximum likelihood (ML) procedure for RBM studied here, namely the existence of (1) localized receptive field focusing on a subset of strongly correlated inputs, and of (2) cross-inhibition between hidden units during the learning phase, which makes their place/receptive fields repel each other and forces them to cover as much as possible of the input space (figure 4). Achieving a more precise understanding of how general this scenario is, and how it extends to deeper architectures i.e with more neural layers would be very interesting.

Acknowledgments

This work was partly funded by the ANR project RBMPro CE30-0021-01 and the HFSP project RGP0057-2016. MH benefited from a fellowship from the ICFP Labex of Department of Physics at ENS. JT acknowledges partial support by a fellowship from the Edmond J Safran Center for Bioinformatics at Tel Aviv University.

Appendix. Conditions for pattern formation

The continuous partial differential equation (38) along with (39) describes the evolution of the field $w(x, t)$ in space and time. This equation will only lead to a non-trivial stable steady state pattern if certain conditions that we make explicit below are fulfilled.

We start from (39) and differentiate this equation with respect to time (setting $\nu = 1$ to lighten notations) to get

$$\frac{db(t)}{dt} = 4b(t) - 2b(t)^2 + 2 \int_0^L w(x, t) \frac{\partial^2}{\partial x^2} w(x, t) dx + 2 \int_0^L w^4(x, t) dx. \quad (\text{A1})$$

The non-trivial, uniform fixed points of this equation is

$$w^* = \sqrt{\frac{2}{L-1}} \quad \text{and} \quad b^* = \frac{2L}{L-1}. \quad (\text{A2})$$

One can rewrite the above equations in the following simple notation:

$$\begin{aligned} \frac{\partial w}{\partial t} &= \gamma f(w, b) + \Delta w \\ \frac{db}{dt} &= \gamma g(w, b) + d\Delta b \end{aligned} \quad (\text{A3})$$

where d is eventually sent to infinity, since there is no spatial time lag for reaching the equilibrium value of b , and b is always spatially uniform. After linearization around the fixed point for small $|\mathbf{w}|$:

$$\mathbf{w} = \begin{pmatrix} w - w^* \\ b - b^* \end{pmatrix},$$

these two equations can be written in vector form as follows

$$\begin{aligned} \frac{\partial \mathbf{w}}{\partial t} &= \gamma A \mathbf{w} + D \Delta \mathbf{w}, \quad \text{where } D = \begin{pmatrix} 1 & 0 \\ 0 & d \end{pmatrix} \quad \text{and} \\ A &= \begin{pmatrix} f_w & f_b \\ g_w & g_b \end{pmatrix}_{(w^*, b^*)} = \begin{pmatrix} 2 + 3(w^*)^2 - b^* & -w^* \\ 8(w^*)^3 L & 4(1 - b^*) \end{pmatrix} \end{aligned} \quad (\text{A4})$$

are, respectively, the diffusion and stability matrix.

We impose first that the uniform fixed point should be stable in the absence of any spatial variation, as we demand that the instability solely comes from spatial interactions. Keeping the non-spatial part of the equation:

$$\frac{\partial \mathbf{w}}{\partial t} = \gamma A \mathbf{w} \quad (\text{A5})$$

We look for solutions of the form $\mathbf{w} = e^{\lambda t} \mathbf{w}_0$. Stability requires that $\text{Re}(\lambda) < 0$, that is,

$$\begin{aligned} \text{tr } A &= f_w + g_b < 0 \\ \det A &= f_w g_b - f_b g_w > 0 \end{aligned} \quad (\text{A6})$$

It is easy to check that these general conditions, once applied to the derivatives of f and g listed in (A4), are satisfied as soon as $L > 1$.

We then ask for the existence of an instability resulting from the spatial part of the equation. As the Laplacian operator is translation invariant we look for a solution to the reaction–diffusion system (A4) that can be decomposed onto Fourier wave planes of momentum k multiple of $2\pi/L$ due to periodic boundary conditions:

$$\mathbf{w}(x, t) = \sum_k c_k e^{\lambda t} \mathbf{W}_k(x), \quad (\text{A7})$$

where the constants c_k are determined by a Fourier expansion of the initial conditions in terms of $\mathbf{W}_k(x)$ and λ is the eigenvalue that determines the temporal growth of the instability. Inserting (A7) into (A4), we get for each k ,

$$\lambda \mathbf{W}_k = \gamma A \mathbf{W}_k - k^2 D \mathbf{W}_k. \quad (\text{A8})$$

Hence, λ is the root of the following characteristic polynomial:

$$\det(\lambda I - \gamma A + k^2 D) = 0. \quad (\text{A9})$$

For the uniform steady state (w^*, b^*) to be unstable against spatial fluctuations, we require $\text{Re}(\lambda) > 0$ for some $k \neq 0$. The conditions for this can be easily worked out, with the result

$$df_w + g_b > 0, \quad (\text{A10})$$

and

$$(df_w + g_b)^2 - 4d(f_w g_b - f_b g_w) > 0. \quad (\text{A11})$$

Condition (A11) is always satisfied since $d \rightarrow +\infty$. To check (A10) we have to evaluate the coefficient of d , which is f_w :

$$f_w = 2 + 3(w^*)^2 - b^* = \frac{4}{L-1} > 0. \quad (\text{A12})$$

Hence, this condition is satisfied as soon as $L > 1$.

ORCID iDs

Rémi Monasson  <https://orcid.org/0000-0002-4459-0204>

References

- [1] Bengio Y, Courville A and Vincent P 2013 Representation learning: a review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1798–828
- [2] Laughlin S 1981 A simple coding procedure enhances a neuron's information capacity *Z. Naturforsch.* **36** 910–2
- [3] Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Netw., Comput. Neural Syst.* **3** 213–51
- [4] Garrigan P, Ratliff C P, Klein J M, Sterling P, Brainard D H and Balasubramanian V 2010 Design of a trichromatic cone array *PLoS Comput. Biol.* **6** e1000677
- [5] Tesileanu T, Cocco S, Monasson R and Balasubramanian V 2019 Adaptation of olfactory receptor abundances for efficient coding *eLife* **8** e39279
- [6] Brito C S N and Gerstner W 2016 Nonlinear Hebbian learning as a unifying principle in receptive field formation *PLoS Comput. Biol.* **12** e1005070
- [7] Olshausen B A and Field D J 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9
- [8] Hyvärinen A and Oja E 2000 Independent component analysis: algorithms and applications *Neural Network.* **13** 411–30
- [9] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257
- [10] Monasson R 1993 Storage of spatially correlated patterns in auto-associative memories *J. Phys. I* **3** 1141
- [11] Sengupta A M, Tepper M, Pehlevan C, Genkin A and Chklovskii D B 2018 Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks *Adv. Neural Inf. Process. Syst.* **31** 7080–90

- [12] Benna M K and Fusi S 2019 Are place cells just memory cells? Memory compression leads to spatial tuning and history dependence (bioRxiv:[624239](#))
- [13] Battista A and Monasson R 2020 Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks *Phys. Rev. Lett.* **124** 048302
- [14] Mehta P and Schwab D J 2014 An exact mapping between the Variational Renormalization Group and Deep Learning (arXiv:[1410.3831](#))
- [15] Koch-Janusz M and Ringel Z 2018 Mutual information, neural networks and the renormalization group *Nat. Phys.* **14** 578–82
- [16] Watkin T L H and Nadal J-P 1994 Optimal unsupervised learning *J. Phys. A: Math. Gen.* **27** 1899
- [17] Reimann P and Van den Broeck C 1996 Learning by examples from a nonuniform distribution *Phys. Rev. E* **53** 3989–98
- [18] Baik J, Ben Arous G and Peché S 2005 Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices *Ann. Probab.* **33** 1643–97
- [19] Barra A, Genovese G, Sollich P and Tantari D 2017 Phase transitions in Restricted Boltzmann Machines with generic priors *Phys. Rev. E* **96** 042156
- [20] Le Roux N and Bengio Y 2008 Representational power of restricted Boltzmann machines and deep belief networks *Neural Comput.* **20** 1631–49
- [21] Tubiana J and Monasson R 2017 Emergence of compositional representations in restricted Boltzmann machines *Phys. Rev. Lett.* **118** 138301
- [22] Tieleman T 2008 Training restricted Boltzmann machines using approximations to the likelihood gradient *Proc. 25th Int. Conf. on Machine Learning* pp 1064–71
- [23] Fischer A and Igel C 2012 An introduction to restricted Boltzmann machines *Iberoamerican Congress on Pattern Recognition* (Berlin: Springer) pp 14–36
- [24] Jastrzebski S, Kenton Z, Arpit D, Ballas N, Fischer A, Bengio Y and Storkey A 2017 Three factors influencing minima in SGD (arXiv:[1711.04623](#))
- [25] Hochreiter S and Schmidhuber J 1997 Flat minima *Neural Comput.* **9** 1–42
- [26] Keskar N S, Mudigere D, Nocedal J, Smelyanskiy M and Tang P T P 2016 On large-batch training for deep learning: generalization gap and sharp minima (arXiv:[1609.04836](#))
- [27] Chaudhari P, Choromanska A, Soatto S, LeCun Y, Baldassi C, Borgs C, Chayes J, Sagun L and Zecchina R 2016 Entropy-SGD: biasing gradient descent into wide valleys (arXiv:[1611.01838](#))
- [28] Bottou L and Bousquet O 2008 The tradeoffs of large scale learning *Adv. Neural Inf. Process. Syst.* **20** 161–8
- [29] Smith S L and Le Q V 2017 A Bayesian perspective on generalization and stochastic gradient descent (arXiv:[1710.06451](#))
- [30] Marčenko V A and Pastur L A 1967 Distribution of eigenvalues for some sets of random matrices *Math. USSR Sbornik* **1** 457
- [31] Decelle A, Fissore G and Furtlehner C 2017 Spectral dynamics of learning in restricted Boltzmann machines *Europhys. Lett.* **119** 60001
- [32] Murray J D 2003 *Mathematical Biology. II. Spatial Models and Biomedical Applications (Interdisciplinary Applied Mathematics vol 18)* (New York: Springer) (doi:[10.1007/b98869](#))
- [33] Amari S 1977 Dynamics of pattern formation in lateral-inhibition type neural fields *Bio. Cyber.* **27** 77–87
- [34] Tsodyks M and Sejnowski T 1995 Associative memory and hippocampal place cells *Int. J. Neur. Syst.* **6** 81–6
- [35] Ben-Yishai B, Bar-Or R and Sompolinsky H 1995 Theory of orientation tuning in visual cortex *Proc. Natl. Acad. Sci.* **92** 3844–8
- [36] Treves A 2019 private communication
- [37] Casali G, Bush D and Jeffery K 2019 Altered neural odometry in the vertical dimension *Proc. Natl. Acad. Sci.* **116** 4632–6
- [38] Hayman R, Verriotes M, Jovalekic A, Fenton A A and Jeffery K 2011 Anisotropic encoding of three-dimensional space by place cells and grid cells *Nat. Neurosci.* **14** 1182–8
- [39] Chen G, King J A, Burgess N and O’Keefe J 2010 How vision and movement combine in the hippocampal place code *Proc. Natl. Acad. Sci.* **110** 378–83