

Emergence of Compositional Representations in Restricted Boltzmann Machines

J. Tubiana, R. Monasson

*Laboratoire de Physique Théorique, Ecole Normale Supérieure and CNRS,
PSL Research, Sorbonne Universités UPMC, 24 rue Lhomond, 75005 Paris, France*

(Dated: March 6, 2017)

Extracting automatically the complex set of features composing real high-dimensional data is crucial for achieving high performance in machine-learning tasks. Restricted Boltzmann Machines (RBM) are empirically known to be efficient for this purpose, and to be able to generate distributed and graded representations of the data. We characterize the structural conditions (sparsity of the weights, low effective temperature, nonlinearities in the activation functions of hidden units, and adaptation of fields maintaining the activity in the visible layer) allowing RBM to operate in such a compositional phase. Evidence is provided by the replica analysis of an adequate statistical ensemble of random RBMs and by RBM trained on the handwritten digits dataset MNIST.

Recent years have witnessed major progresses in supervised machine learning, *e.g.* in video, audio, image processing,... [1]. Despite those impressive successes, unsupervised learning, in which the structure of data is learned without *a priori* knowledge still presents formidable challenges. A fundamental question is how to learn probability distributions that fit well complex data manifolds in high-dimensional spaces [2]. Once learnt, such *generative models* can be used for denoising, completion, artificial data generation,... Hereafter we focus on one important generative model, Restricted Boltzmann Machines (RBM) [3, 4]. In its simplest formulation a RBM is a Boltzmann machine on a bipartite graph, see Fig. 1(a), with a visible (v_i) layer that represents the data, connected to a hidden (h) layer meant to extract and explain their statistical features. The marginal distribution over the visible layer is fitted to the data through approximate likelihood maximization [5–8]. Once the parameters are trained each hidden unit becomes selectively activated by a specific data feature; owe to the bidirectionality of connections the probability to generate configurations of the visible layer where this feature is present is, in turn, increased. Multiple combinations of numbers of features, with varying degrees of activation of the corresponding hidden units allow for efficient generation of a large variety of new data samples. However, the existence of such ‘compositional’ encoding seems to depend on the values of the RBM parameters, such as the size of the hidden layer [9]. Characterizing the conditions under which RBM can operate in this compositional regime is the purpose of the present work.

In the RBM shown in Fig. 1(a) the visible layer includes N units v_i , with $i = 1, \dots, N$, chosen here to be binary ($= 0, 1$). Visible units are connected to M hidden units h_μ , through the weights $\{w_{i\mu}\}$. The energy of a configuration $\mathbf{v} = \{v_i\}$, $\mathbf{h} = \{h_\mu\}$ is defined through

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

where \mathcal{U}_μ is a potential acting on hidden unit μ ; due

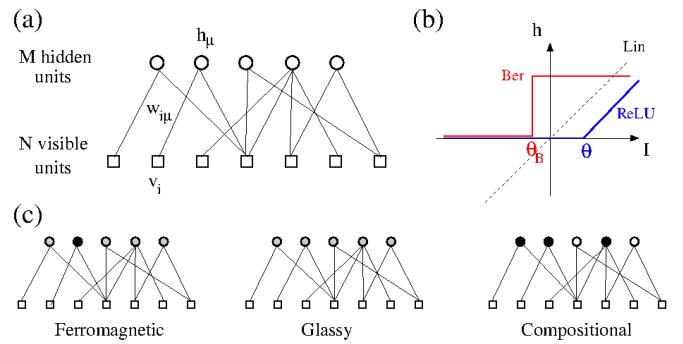


FIG. 1: (a) Architecture of RBM. Visible (v_i , $i = 1, \dots, N$) and hidden (h_μ , $\mu = 1, \dots, M$) units are connected through weights ($w_{i\mu}$). (b) Activation functions Φ of Bernoulli, Linear and Rectified Linear Units. The corresponding potentials are $\mathcal{U}^{Lin}(h) = \frac{h^2}{2}$; $\mathcal{U}^{Ber}(h) = h \theta_B$ if $h = 0$ or 1, and $+\infty$ otherwise; $\mathcal{U}^{ReLU}(h) = \frac{h^2}{2} + h \theta$ for $h \geq 0$, $+\infty$ for $h < 0$. (c) The three regimes of operation, see text. Black, grey and white hidden units symbolize, respectively, strong, weak and null activations.

to the binary nature of the visible units their potential is fully characterized by a local field, g_i in (1). Configurations are then sampled from the Gibbs equilibrium distribution associated to E , $P[\mathbf{v}, \mathbf{h}] = \exp(-E[\mathbf{v}, \mathbf{h}])/Z$, where Z is the partition function [3].

Given a visible configuration \mathbf{v} the most likely value h_μ of hidden unity μ is a function of its input $I_\mu = \sum_{i=1}^N w_{i\mu} v_i$: $h_\mu = \Phi_\mu(I_\mu)$, where the activation function $\Phi_\mu = (\mathcal{U}'_\mu)^{-1}$ as can be seen from the minimization of E . Examples of Φ are shown in Fig. 1(b). When Φ is linear, *i.e.* for quadratic potential \mathcal{U} the probability $P[\mathbf{v}, \mathbf{h}]$ is Gaussian in the hidden units, and the marginal distribution $P[\mathbf{v}]$ of the visible configurations \mathbf{v} can be exactly computed [10]. It coincides with the equilibrium distribution of a Boltzmann machine with a pairwise interaction matrix $J_{ij} = \sum_\mu w_{i\mu} w_{j\mu}$, or, equivalently, of a Hopfield model [11], whose M patterns \mathbf{w}_μ are the columns of the weight matrix $\{w_{i\mu}\}$.

Activation functions Φ empirically known in machine-

learning literature to provide good results are, however, nonlinear. Nonlinear Φ produce effective Boltzmann machines with high order (> 2) multibody interactions between the visible units v_i . Two examples are shown in Fig. 1(b): Bernoulli units, which take discrete 0,1 values, and Rectified Linear Units (ReLU) [1]. Unlike Bernoulli units ReLU preserve information about the magnitudes of their inputs above threshold [12]; this property is expected for real neurons and ReLU were first introduced in the context of theoretical neuroscience [13].

We first report results from a training experiment of RBM with ReLU on the handwritten digits dataset MNIST [14]. Our goal is not to classify digits from 0 to 9, but to learn a generative model of digits from examples. Details about learning can be found in Supplemental Material, Section I. Figure 2(a) shows typical ‘features’ $w_\mu = \{w_{i\mu}\}$ after learning. Each feature includes negative and positive weights, and is localized around small portions of the visible layer. These features look like elementary strokes, which are combined by the RBM to generate random digits (Fig. 2(b)). In each generated handwritten digit image $\hat{S} \simeq 240$ hidden units are silent ($h_\mu = 0$), see histogram in Fig. 2(c). The remaining hidden units have largely varying activations, some weak and few very strong; we estimate the number of strongly activated ones through the participation ratio $\hat{L} = [(\sum_\mu h_\mu^a)^2 / \sum_\mu h_\mu^{2a}]$, with exponent $a = 3$ as explained below. On average $\hat{L} \simeq 20$ elementary strokes compose a generated digit, see Fig. 2(c). Different combinations of strokes correspond to different variants of the same digits. Many of those variants are not contained in the training set, and closely match digits in the test set (Supplemental Material, Fig. 1(b)), hence showing the generative power of RBM.

Learning is accompanied by structural changes in RBM, which we track with two parameters: $\hat{p} = \frac{1}{MN} \sum_\mu [(\sum_i w_{i\mu}^2)^2 / \sum_i w_{i\mu}^4]$ and $W_2 = \frac{1}{M} \sum_{i,\mu} w_{i\mu}^2$. Those parameters are proxies for, respectively, the fraction of nonzero weights and the effective inverse temperature, see Supplemental Material, Section III. Figure 2(d) shows that \hat{p} diminishes to small values ~ 0.1 , whereas W_2 increases. While most weights become very small and negligible the remaining ones get large, in agreement with Fig. 2(a). Notice that sparsity is not imposed to obtain a specific class of features, *e.g.* as in [15], but naturally emerges through likelihood maximization across training. The presence of large weights implies that flipping visible units is associated to large energy costs. Visible units are effectively at very low temperature, as can be seen from the quasi binary nature of conditional averages in Fig. 2(b) and Supplemental Material, Fig. 5.

We argue below that these structural changes are not specific to MNIST-trained RBM but are generically needed to bring RBM towards a compositional phase, in which visible configurations are composed from combina-

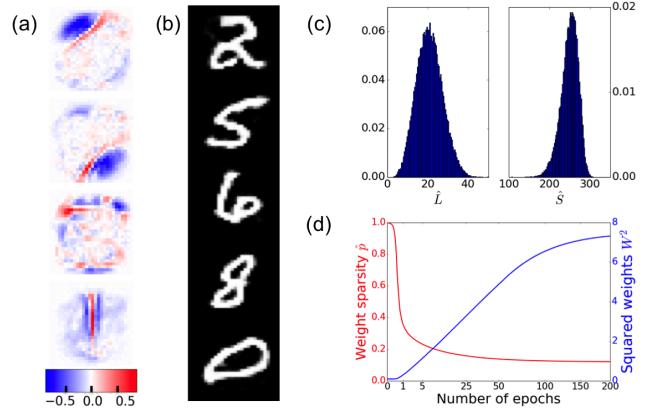


FIG. 2: Training of RBM on MNIST, with $N = 28 \times 28$ visible units and $M = 400$ hidden ReLU. (a) Set of weights w_μ attached to four hidden units μ . (b) Averages of v conditioned to five hidden-unit configurations h sampled from the RBM at equilibrium. Black and white pixels correspond respectively to averages equal to 0 and 1; few intermediary values, indicated by grey levels, can be seen on the edges of digits. (c) Distributions of \hat{L} (left) and \hat{S} (right) in hidden-unit configurations at equilibrium. (d) Evolution of the weight sparsity \hat{p} (red) and the squared weight value W_2 (blue); The training time is measured in epochs (number of passes over the data set), and represented on a square-root scale.

tions of a large number L (typically, $1 \ll L \ll M$, as in Fig. 2(c)) of features encoded by simultaneously, strongly activated hidden units. Our claim is supported by a detailed analysis of a Random RBM (R-RBM) ensemble, in which the weights $w_{i\mu}$ are quenched random variables, with controlled sparsity and strength, and the magnitude of the visible fields and the values of the ReLU thresholds can be chosen. For adequate choices of these control parameters the compositional phase is thermodynamically favoured with respect to the ferromagnetic phase of the Hopfield model, where one pattern is activated [16], and to the spin-glass phase, in which all hidden units are weakly and incoherently activated (Fig. 1(c)).

In the R-RBM ensemble weights $w_{i\mu}$ are independent random variables, equal to $-\frac{1}{\sqrt{N}}, 0, +\frac{1}{\sqrt{N}}$ with probabilities equal to, respectively, $\frac{p_i}{2}, 1-p_i, \frac{p_i}{2}$; $p_i \in [0; 1]$ sets the degree of sparsity of the weights attached to the visible unit v_i , high sparsities corresponding to small p_i . The estimator \hat{p} defined above (Fig. 2(d)) measures the fraction of nonzero weights, $p = \sum_i p_i/N$. This distribution was previously introduced to study parallel storage of multiple sparse items in the Hopfield model [17, 18]. For simplicity the fields on visible units and the potentials acting on hidden units are chosen to be uniform, $g_i = g$ and $\mathcal{U}_\mu = \mathcal{U}^{ReLU}$ (Fig. 1(a)). We define the ratio of the numbers of hidden and visible units, $\alpha = M/N$.

Given a visible layer configuration v , hidden units μ coding for features w_μ present in v will be strongly ac-

tivated: their inputs $I_\mu = \mathbf{w}_\mu \cdot \mathbf{v}$ will be strong and positive, comparable to the product of the norms of \mathbf{w}_μ ($\simeq \sqrt{p}$ for large N) and \mathbf{v} (of the order of \sqrt{pN}), and, hence, will scale as $m\sqrt{N}$, where m , called magnetization, is finite. Most hidden units μ' have, however, features $\mathbf{w}_{\mu'}$ essentially orthogonal to \mathbf{v} , and receive inputs $I_{\mu'}$ fluctuating around 0, with finite variances. These scalings ensure that \hat{L} defined above (Fig. 2(c)) will coincide with the number L of strongly activated units when $N \rightarrow \infty$; choosing exponent $a = 2$ in \hat{L} rather than $a = 3$ would have introduced biases coming from weakly activated units (Supplemental Material, Section III.B).

The typical ground state (GS) energy E (1) of R-RBM can be computed with the replica method within the replica-symmetric Ansatz [16], as the optimum of

$$\begin{aligned} E_{GS} &= \frac{L}{2}m^2 + \frac{\alpha}{2}(qB + rC) - \frac{1}{N} \sum_i \sqrt{\alpha p_i r} \quad (2) \\ &\times \left\langle H^{(1)} \left(- \left[g + \frac{\alpha}{2}Bp_i + mW \right] / \sqrt{\alpha p_i r} \right) \right\rangle_W \\ &+ \alpha \int Dz \min_h \left(\mathcal{U}^{ReLU}(h) - \frac{C}{2}h^2 - z\sqrt{pq}h \right) \end{aligned}$$

over the order parameters m, L, r, q, B, C (averaged over the quenched weights): m and L are, respectively, the magnetization and the number of feature-encoding hidden units, r is the mean squared activity of the other hidden units, $q = \sum_i p_i \langle v_i \rangle_{GS} / (Np)$ is the weighted activity of the visible layer in the GS, and B, C are response functions, *i.e.* derivatives of the mean activity of, respectively, hidden and visible units with respect to their inputs [20]. In (2) $Dz = \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}$ denotes the Gaussian measure, $H^{(k)}(x) = \int_x Dz(z-x)^k$, and $\langle \cdot \rangle_W$ is the average over the sum W of L i.i.d. weights $w_{i\mu}$ drawn as above.

We first fix L , and optimize E_{GS} over all the other order parameters. At large α the only solution has $m = 0$, and corresponds to the Spin-Glass phase. For intermediate values of α , other solutions, with $m > 0$, exist. For the sake of simplicity we consider first the homogenous sparsity case, with $p_i = p$. We show in Fig. 3(a), for fixed $p = 0.1$ and various values of L , the maximal value of α below which a phase with L magnetized hidden units exists. Importantly this critical value can be made arbitrarily large by increasing the ReLU threshold θ . This phenomenon is a consequence of the nonlinearity of ReLU, and can be understood as follows. The squared activity r of non-magnetized hidden units obeys the saddle-point equation $r = pq/(1-C)^2 \times H^{(2)}(\theta/\sqrt{pq})$. The first factor is reminiscent of the expression $r = 1/(1-C)^2$ arising for the Hopfield model (for which $p = q = 1$ at zero temperature) [16], while the second factor comes from the nonlinearity of ReLU. $H^{(2)}$ being a rapidly decaying function of its argument large thresholds θ lead to small r values. As the level of crosstalk due to nonmagnetized

hidden units diminishes larger ratios α can be supported by R-RBM without entering the glassy phase. Numerical simulations of R-RBMs at large α confirm the existence and (meta)stability of phases with L nonzero magnetizations (Fig. 3(b)). Moreover, the values of the average normalized magnetizations $\tilde{m} = m/(p/2) \in [0; 1]$ are in excellent agreement with those found by optimizing E_{GS} .

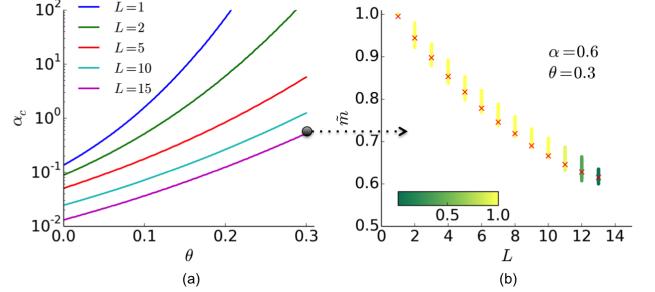


FIG. 3: Compositional regime in R-RBM. (a) Critical lines in the θ, α plane below which L hidden units may be strongly activated, calculated from the optimization of E_{GS} (2). Parameters are $p_i = p = 0.1$, $g = -0.02$. (b) Comparison of theoretical (red crosses) and numerical simulations ($N = 10^4$ visible units, colored points) for the rescaled magnetizations $\tilde{m} = m/(p/2)$ as a function of the number L of strongly activated hidden units in R-RBM. 7,500 zero temperature MCMC, each initialized with a visible configuration strongly overlapping with $L = 1, 2, \dots$ ‘features’, were launched; color code indicates the probability that the same L hidden units are magnetized after convergence (see bottom scale), and the corresponding average magnetization \tilde{m} .

The nature of the large- L phases and the selection of the value of L are best understood in the limit case of highly sparse connections, $p \ll 1$. The R-RBM model exhibits an interesting limit behaviour, which we call hereafter *compositional* phase. In this regime the number of strongly magnetized hidden units is unbounded, and diverges as $L = \ell/p$, with $\ell > 0$ and finite. The normalized GS energy $e_\ell = E_{GS}(L = \ell/p)/p$ is a nonmonotonic function of the index ℓ , see Fig. 4(a). Minimization of e_ℓ leads to the selection of a well defined index ℓ^* . The magnetizations of the ℓ^*/p strongly activated units, $m = \frac{p}{2}\tilde{m}$, vanish linearly with p [26]. Nonmagnetized hidden units have activities of the order of $\sqrt{r} \sim \sqrt{p}$, and can be shutdown by choosing thresholds $\theta \sim \sqrt{p}$; hence crosstalk between those units can be suppressed, allowing for large relative size α of the hidden layer. The input received by a visible unit from the large number of magnetized units is, after transmission through the dilute weights, of the order of $Lmp = \frac{1}{2}\ell^*\tilde{m}p$; it can be modulated by a (positive or negative) field $g \sim p$ to produce any finite activity q in the visible layer, as soon as the effective temperature gets below $\sim p$.

The compositional phase competes with the ferromagnetic phase, in which $\tilde{m} > 0$ but e_ℓ is a monotonously

growing function of ℓ (hence, $\ell^* = 0$), and the spin glass phase, in which $\tilde{m} = 0$ and e_ℓ does not depend on ℓ , see Fig. 4(a). The phase diagram in the parameter space $(\alpha, \tilde{g} = \frac{q}{p}, \theta = \frac{\theta}{\sqrt{p}})$ will be detailed in [20]. Briefly speaking, given α , $\tilde{\theta}$ should be large enough (as in Fig. 3) and $|\tilde{g}|$ should be neither too large to penalize the ferromagnetic phase, nor too small to avoid the spin glass regime.

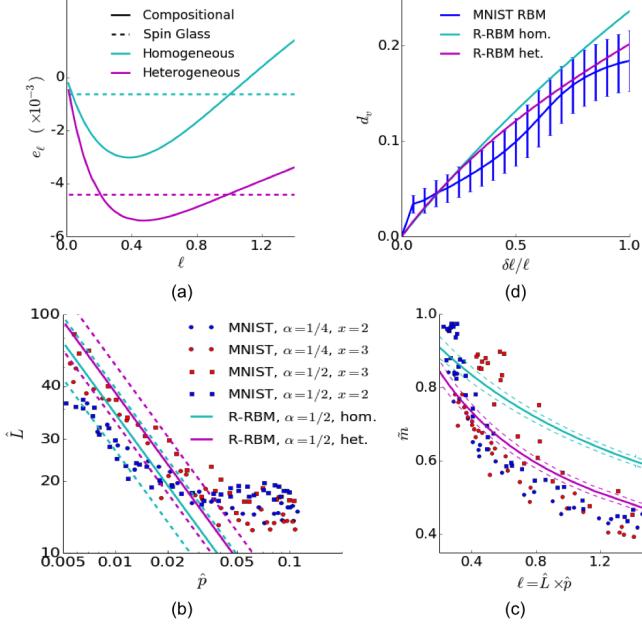


FIG. 4: (a). Behaviour of the GS energy e_ℓ vs. $\ell = L \times p$ in the $p \rightarrow 0$ limit. Parameters: $\tilde{\theta} = 1.5$, $\alpha = 0.5$. (b)-(c)-(d). Quantitative predictions in the compositional regime of R-RBM compared to RBMs inferred on MNIST. Each point represent a ReLU RBM trained with various regularizations, yielding different weight sparsities p . Solid lines depict predictions found by optimizing e_ℓ (2), and dashed line expected fluctuations at finite size (N) and temperature. (b) Average number L of active hidden units vs. p . (c) Average magnetization \hat{m} vs. $\ell = L \times p$. (d) Distance (per pixel) between the pairs of visible configurations after convergence of zero-temperature MCMC vs. relative distances in the hidden-unit activation patterns. MCMC are initialized with all pairs of digits in MNIST; final visible configurations differ from MNIST digits by about 7 pixels, both on the training and test sets, see Supplemental Material, Fig. 1. In all four panels predictions for the homogeneous ($\tilde{g} = -0.21$) and heterogeneous ($\tilde{g} = -0.1725$, see Supplemental Material, Section III.E) cases are shown in, respectively, cyan and magenta.

Characteristic properties of our compositional phase are confronted to ReLU RBMs trained on MNIST in Fig. 4 (b,c). Compared to Fig. 2 we add a regularization penalty $\propto \sum_\mu (\sum_i |w_{i\mu}|)^x$ to control the final degree of sparsity; the case $x = 1$ gives standard L_1 regularization, while, for $x > 1$, the effective penalty strength $\propto (\sum_i |w_{i\mu}|)^{x-1}$ increases with the weights, hence promoting homogeneity among hidden units. After train-

ing we generate Monte Carlo samples of each RBM at equilibrium, and monitor the average number of active hidden units, \hat{L} , and the normalized magnetization, \hat{m} . Figure 4(b) shows \hat{L} vs. \hat{p} , in good agreement with the R-RBM theoretical scaling $L \sim \frac{\ell^*}{p}$. Figure 4(c) shows that \hat{m} is a decreasing function of $\ell = \hat{L} \times \hat{p}$, as qualitatively predicted by theory, but quantitatively differs from the prediction of R-RBM with homogeneous p . This disagreement can be partly explained by the heterogeneities in the sparsities p_i in RBMs trained on MNIST, *e.g.* units on the borders are connected to only few hidden units, whereas units at the center of the grid are connected to many. We introduce a heterogeneous R-RBM model, where the distribution of the p_i 's is fitted from MNIST-trained RBMs (Supplemental Material, Section III.E). Its GS energy can be calculated from (2), see Fig. 4(a) [20]. Results are shown in Fig. 4(b,c) to be in good agreement with RBM trained on MNIST.

RBM, unlike the Hopfield or mixture model, may produce gradually different visible configurations through progressive changes in the hidden-layer activation pattern. R-RBMs enjoy the same property. We compute, through a real-replica approach [20], the average Hamming distance d (per pixel) between the visible configurations $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}$ minimizing the energy E (1) for two hidden configurations $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}$ sharing $(\ell - \delta\ell)/p$ hidden units among the ℓ/p strongly activated ones. Figure 4(d) shows that d monotonously increases from $d = 0$ for $\delta\ell = 0$ up to $d = 2q(1 - q)$ (complete decorrelation of visible units) for $\delta\ell = \ell$, in very good quantitative agreement with results for RBM trained on MNIST.

The gradual change property has deep dynamical consequences. MCMC of MNIST-trained RBM (videos available in Supplemental Material) show that gradual changes may occasionally lead to another digit type, by passing through well-drawn, yet ambiguous digits. The progressive replacement of feature-encoding hidden units (small $\delta\ell$ steps) along the transition path does not increase much the energy, and the transition process is fast compared to activated hopping between deep minima taking place in the Hopfield model.

Our study is related to several previous works. RBMs with linear activation function Φ coincide with the Hopfield model. In this framework magnetized hidden units identify retrieved patterns, and α corresponds to the capacity of the autoassociative memory. Tsodyks and Feigel'man showed how the critical capacity (for single pattern retrieval) could be dramatically increased with sparse weights ($p \ll 1$) and appropriate tuning of the fields g_i [21]; however this effect could be achieved only with vanishingly low activities q . Agliari and collaborators showed in a series of papers [17, 18] that multiple sparse patterns could be simultaneously retrieved in the case of linear Φ and vanishing capacity $\alpha = 0$ (finite M). Finite capacity $\alpha \sim c^{-2}$ could be achieved at zero

temperature in the limit of extreme sparsity, $p = c/N$, only [19]; for MNIST $p \simeq 0.1$ and $N = 784$ would give $\alpha \sim 2.10^{-4}$. Our work shows that large values of α can be reached even with moderate sparsity p (as in realistic situations, see Fig. 2) provided that nonlinear Φ (ReLU) and appropriate threshold values θ are considered. The presence of the fields g_i acting on the visible units (absent in the $v_i = \pm 1$ model of [17–19]), is also crucial for the existence of our compositional phase as explained above.

It would be interesting to extend our work to more than one layers of hidden units, or to other types of nonlinear Φ . While numerical studies of RBMs with Bernoulli hidden units show no qualitative change compared to ReLU, choosing $\Phi(h)$ growing asymptotically faster than h could affect the nature of the extracted features [23]. An important challenge would be to understand the training dynamics, *i.e.* how hidden units gradually extract features from data prototypes.

Acknowledgements. We are grateful to C. Fisher and G. Semerjian for useful discussions. This work was partly funded by the CNRS-Inphyniti Inferneuro and the HFSP RGP0057/2016 projects, and benefited from the support of NVIDIA Corporation with the donation of a Tesla K40 GPU card.

- [1] Y. LeCun, Y. Bengio, J. Hinton, *Nature* **521**, 436-444 (2015).
- [2] Y. Bengio, A. Courville, P. Vincent, *IEEE transactions on pattern analysis and machine intelligence* **35**, 1798-1828 (2013).
- [3] P. Smolensky, Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, MIT Press, 194-281 (1986).
- [4] R. Salakhutdinov, A. Mnih, G. Hinton, *Proceedings of the 24th international conference on Machine learning*, p. 791-798 (2007).
- [5] G. Hinton, *Momentum* **9**, 926 (2010).
- [6] T. Tieleman, *Proceedings of the 25th international conference on Machine learning*, pp. 1064-1071 (2008).
- [7] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, O. Delalleau, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 145-152 (2010).
- [8] M. Gabrie, E.W. Tramel, F. Krzakala, *Advances in Neural Information Processing Systems*, pp. 640-648 (2015).
- [9] A. Fischer, C. Igel, *Iberoamerican Congress on Pattern Recognition*, pp. 14-36 (2012).
- [10] A. Barra, A. Bernacchia, E. Santucci, P. Contucci, *Neural Networks* **34**, 1-9 (2012).
- [11] J.J. Hopfield, *Proc. Nat. Acad. Sci. USA* **79**, 2554 (1982).
- [12] V. Nair, G.E. Hinton, *Proceedings of the 27th International Conference on Machine Learning*, p. 807-814 (2010).
- [13] A. Treves, *J. Comp. Neurosci.* **2**, 259-272 (1995).
- [14] Y. LeCun, C. Cortes, C.J. Burges, The MNIST database of handwritten digits (1998).
- [15] B.A. Olshausen, D.J. Field, *Nature* **381**, 607-609 (1996).
- [16] D.J. Amit, H. Gutfreund, H. Sompolinsky, *Phys. Rev. Lett.* **55**, 1530 (1985).
- [17] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, F. Moauro, *Phys. Rev. Lett.* **109**, 268101 (2012).
- [18] E. Agliari, A. Annibale, A. Barra, A.C.C. Coolen, D. Tantari, *J. Phys. A* **46**, 415003 (2013).
- [19] P. Sollich, D. Tantari, A. Annibale, A. Barra, *Phys. Rev. Lett.* **113**, 238106 (2014).
- [20] J. Tubiana, R. Monasson, *in preparation* (2017).
- [21] M. Tsodyks, M.V. Feigel'man, *Europhys. Lett.* **6**, 101-105 (1989).
- [22] R. Salakhutdinov, I. Murray, *Proceedings of the 25th international conference on Machine learning*, pp. 872-879 (2008).
- [23] D. Krotov, J.J. Hopfield, *arXiv:1606.01164* (2016).
- [24] See Supplemental Material [url], which includes Ref [25].
- [25] Theano Development team Theano Development Team. *arXiv:1605.02688*. (2016).
- [26] Solutions with nonhomogeneous magnetizations m_μ , varying from one strongly activated hidden unit to another, give additional contributions to E_{GS} of the order of p^2 with respect to the homogeneous solution $m_\mu = m$, and do not affect the value of e_ℓ [20].

Supplemental Material

Emergence of Compositional Representations in Restricted Boltzmann Machines

J. Tubiana, R. Monasson

I. TRAINING RBMS ON MNIST

A. Dataset preparation and initial conditions

- In MNIST, each pixel has a value between 0 and 255. We binarize it by thresholding ≥ 128 . The 28×28 binary images are flattened to a $N = 784$ vector with binary values.
- The dataset is split in a training (60,000 instances) and a test (10,000 instances) sets
- The weights $w_{i\mu}$ are randomly initialized at $\pm W$, where $W = \sqrt{\frac{0.1}{N}}$; this choice corresponds to initial temperature and weight sparsity : $T(0) = 10$ and $p(0) = 1$ (see section III).
- The initial field values are $g_i^0 = \log \left[\frac{\langle v_i \rangle^{MNIST}}{1 - \langle v_i \rangle^{MNIST}} \right]$, where $\langle v_i \rangle^{MNIST}$ denotes the average of pixel i over the training data
- For ReLU, the thresholds θ_μ are all initially set to 0

B. Learning algorithms

A RBM is associated to a probability distribution $P[\mathbf{v}, \mathbf{h}] = \frac{e^{-E[\mathbf{v}, \mathbf{h}]}}{Z}$, where the energy E is defined in the main text. The marginal distribution, $P[\mathbf{v}] = \int \prod dh_\mu P[\mathbf{v}, \mathbf{h}]$ is fitted to the data by likelihood maximization. Given data instances $\mathbf{x}^r, r \in \{1, D\}$, the log-likelihood is :

$$\log \mathcal{L}_{\mathbf{W}, \mathbf{g}, \boldsymbol{\theta}} = \frac{1}{D} \sum_{r=1}^D \log [P[\mathbf{x}^r | \mathbf{W}, \mathbf{g}, \boldsymbol{\theta}]] \quad (1)$$

Where \mathbf{W} is the matrix of weights, \mathbf{g} is the vector of visible layer fields and $\boldsymbol{\theta}$ is the vector of hidden units thresholds. Likelihood maximization is implemented by stochastic gradient descent, with the difficulty that extensive Monte Carlo simulations are required to compute the gradient [1, 2]. For the RBM of Fig. 2 in the main text, we used Persistent Contrastive Divergence [3] with

- 20 mini-batch size
- 100 persistent chains
- 1 Gibbs step between each update
- 200 epochs (600 000 updates in total)
- Initial learning rate of $\lambda_i = 5 \cdot 10^{-3}$, decaying geometrically (decay starts after 60 epochs) to $\lambda_f = 5 \cdot 10^{-4}$

PCD is known to be inaccurate toward the end of learning, because the parameters evolve too fast with respect to the mixing rate of the Markov chains. The regularized RBM of main text, Fig. 4 (b,c), were trained with a more efficient algorithm, variant of Adaptive Parallel Tempering [4, 5] with

- 100 mini-batch size
- 100 persistent chains
- 10 replicas

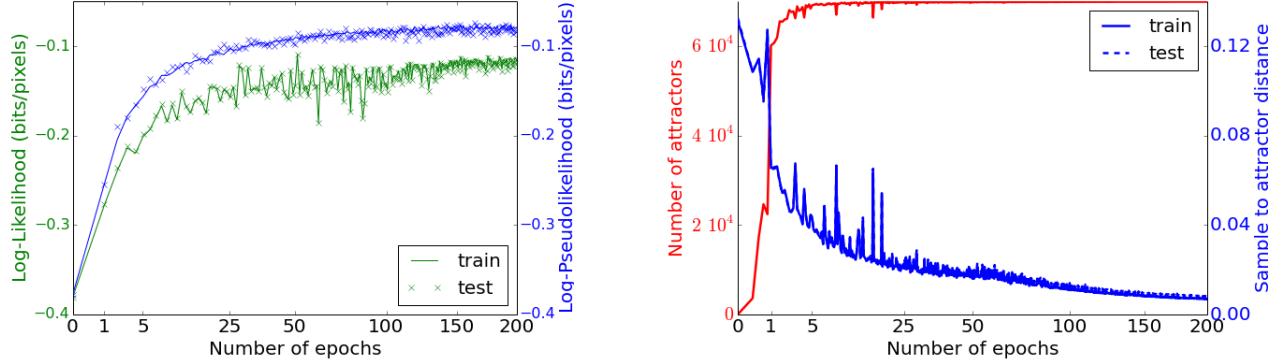


FIG. 1: (a) Evolution throughout training of the data loglikelihood (left scale) and pseudo-loglikelihood (right scale) computed over the training and test sets. (b) Evolution of the number of distinct local maxima of $P_W(v)$ (left scale) and the distance to the original sample (right scale, for train and test set) are displayed.

- 1 Gibbs step between each update
- 150 epochs (90 000 updates in total)
- Initial learning rate of $\lambda_i = 10^{-2}$, decaying geometrically (decay starts after 90 epochs) to $\lambda_f = 10^{-4}$

C. Monitoring the learning

We monitor the evolution of the likelihood and of the pseudo-likelihood of the train and test data sets throughout learning, see Fig. 1(a). The choice of parameters made learning slow, but ensured that the likelihood increased steadily throughout training. The likelihood requires approximate computation of the model partition function; Annealed Importance Sampling [6] was used. Parameters : $n_\beta = 10000$ inverse temperatures with an adaptive spacing [5], $n_{runs} = 1$. Additionnaly, we can look at the probability landscape $P_W(v)$ throughout learning. For each of the 70k MNIST samples, a gradient ascent on $P_W(v)$ is performed until convergence to a local maximum; the number of distinct local maxima of $P_W(v)$ and the distance to the original sample are measured. As training goes, more local maxima appear, and they get closer to the training samples; local maxima also appear close to the test set, which shows that RBM generalize well.

D. Controlling weight sparsity with regularization

To control the weight sparsity p , a regularization penalty is added to the log likelihood $\log \mathcal{L}_{\mathbf{W}, \mathbf{g}, \boldsymbol{\theta}}$:

$$\begin{aligned} \text{Cost} &= -\log \mathcal{L}_{\mathbf{W}, \mathbf{g}, \boldsymbol{\theta}} + L^{(x)} \\ L^{(x)} &= \frac{\lambda_x}{x} \sum_{\mu} \left[\sum_i |w_{i\mu}| \right]^x \\ -\frac{\partial \text{Cost}}{\partial w_{i\mu}} &= \frac{\partial}{\partial w_{i\mu}} \log \mathcal{L}_{\mathbf{W}, \mathbf{g}, \boldsymbol{\theta}} - \lambda_x \left[\sum_j |w_{j\mu}| \right]^{x-1} \text{sign}(w_{i\mu}) \end{aligned} \quad (2)$$

The case $x = 1$ is the usual L_1 penalty and performing gradient descent with $\lambda_1 > 0$ is known to reduce the number of non-zero weights $w_{i\mu}$. However, experiments show that the outcome is inhomogeneous with respect to the hidden units: some hidden units are weakly affected by the penalty, whereas some end up completely disconnected from the visible layer, making them useless, see Fig. 2. To maintain homogeneity among the hidden units, we pick $x = 2$ or $x = 3$. As can be seen from the expression of the gradient, it is equivalent to a usual L_1 penalty, but with a decay rate adaptive to each hidden unit: hidden units strongly (resp. weakly) coupled to the visible layer (large $\sum_i |w_{i\mu}|$) are strongly (resp. weakly) regularized, thus increasing the homogeneity among hidden units.

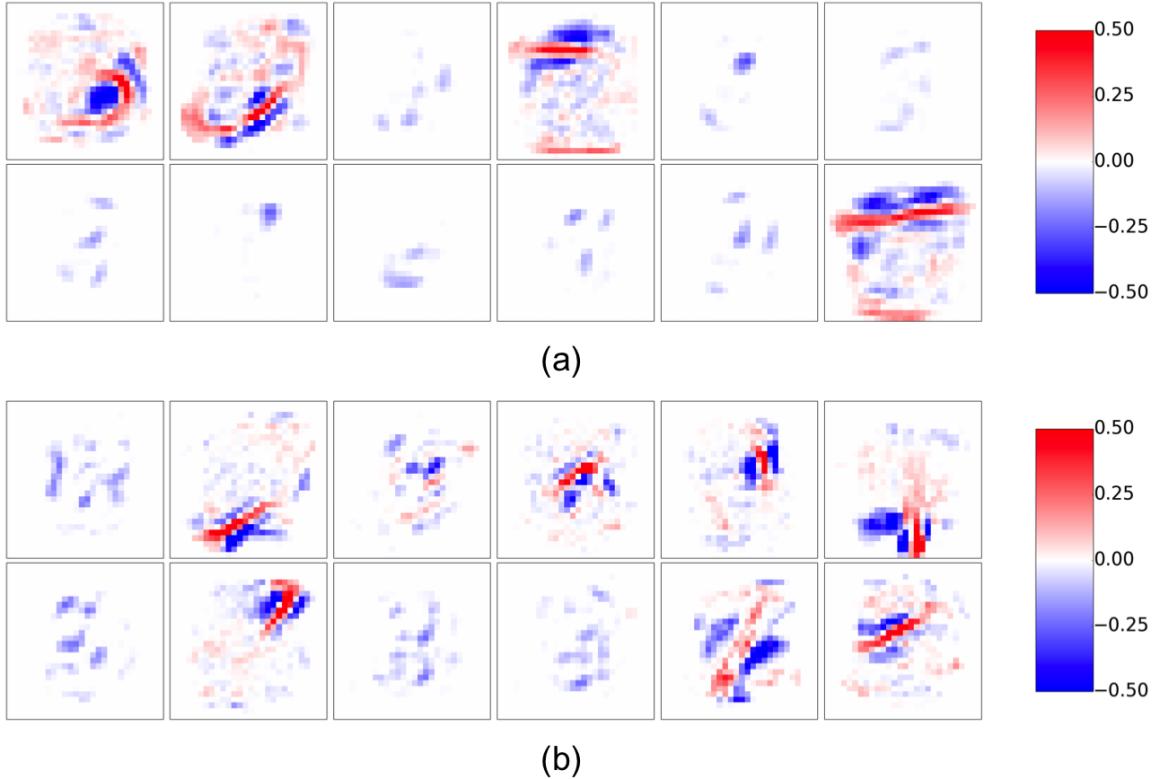


FIG. 2: Subset of 12 weight features produced by training on MNIST, regularized with $L^1, \lambda_1 = 10^{-3}$ (top panel), and $L^2, \lambda_2 = 3.10^{-5}$ (bottom panel). Both have overall sparsity $p \sim 0.036$, but the latter is more homogeneously distributed across hidden units.



FIG. 3: Six independent Monte Carlo Markov Chains realization for a RBM trained on MNIST, extracted from the attached videos, see text.

II. SAMPLING FROM RBMS

RBM can be sampled by Markov Chains Monte Carlo. Due to the conditional independence property, Gibbs sampling can be performed by alternative sampling of \mathbf{h} from $P[\mathbf{h}|\mathbf{v}]$, then \mathbf{v} from $P[\mathbf{v}|\mathbf{h}]$ [1, 2].

A. MCMC Videos

The two videos in Supplementary Material present visualize MCMC runs from RBM trained on MNIST with Bernoulli, Gaussian, ReLU hidden units. Each square depicts a Markov chain started from a random initial condition. 20 Gibbs steps are performed between each image, and each chain is 500 images long. See Fig. 3 for a snapshot.

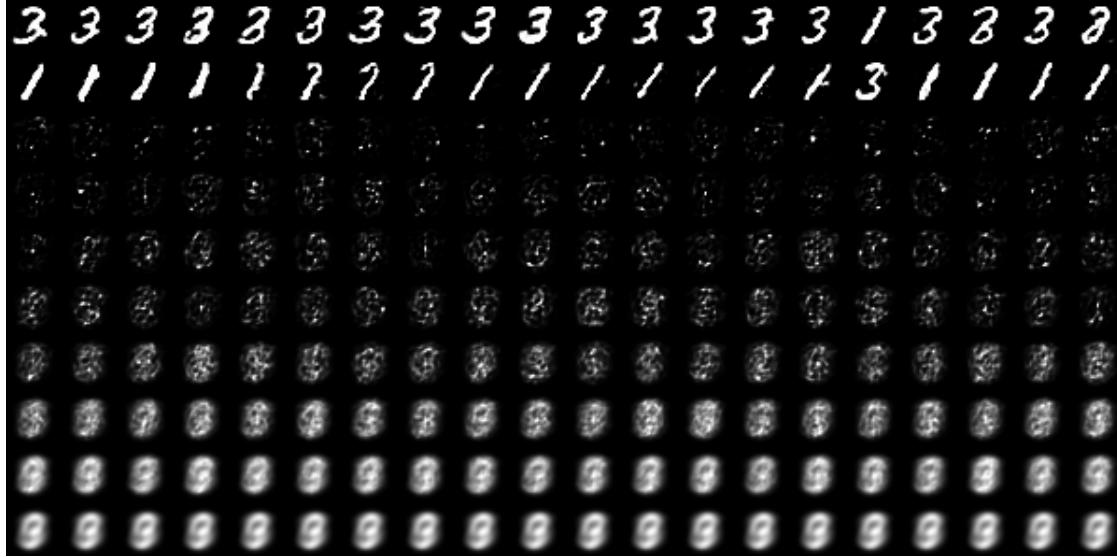


FIG. 4: Ten Monte Carlo Markov Chains realizations at different inverse temperatures, coupled by replica exchange. The plots shows the conditional expectations of visible units, $\mathbb{E}[\mathbf{v}|\mathbf{h}]$, for thermalized hidden-unit activities, \mathbf{h} .

B. Estimating thermal averages with MCMC

Sampling at thermal equilibrium is required to estimate the values of order parameters (L , S , q , \tilde{m} ,...). Since RBM trained on MNIST effectively operate at low temperature (entropy of 0.1 bits/pixel) the MCMC mixing rate is poor, and long simulations would be required for each of the ~ 100 RBMs trained. To overcome this issue we use an Adaptive Parallel Tempering (also known as Replica Exchange) sampling algorithm, with 10 replicas [4, 5]. Observables are averaged over 100 independent Markov Chains, each being first thermalized for 500 Gibbs updates, then run for another 100 Gibbs updates (10K samples in total).

C. Estimating order parameters of R-RBM with zero temperature MCMC

R-RBM are studied analytically in the zero-temperature limit; this limit can be simulated as well. The energy $E[\mathbf{v}, \mathbf{h}]$ of a configuration \mathbf{v}, \mathbf{h} is given by Eqn. (1) in main text, and defines the Gibbs distribution $P^\beta[\mathbf{v}, \mathbf{h}] = \exp(-\beta E[\mathbf{v}, \mathbf{h}]/Z(\beta))$, where $\beta = \frac{1}{T}$ is the inverse temperature. As β increases, $P^\beta[\mathbf{v}, \mathbf{h}]$ is more and more peaked around the minimum of E . In the limit $\beta \rightarrow \infty$, a dynamical Gibbs step becomes deterministic :

$$\begin{aligned} h_\mu &\leftarrow \arg \min_h \left[U_\mu(h) - h \sum_i w_{\mu i} v_i \right] \equiv \Phi_\mu \left[\sum_i w_{i\mu} v_i \right] \\ v_i &\leftarrow \arg \min_v \left[-g_i v - v \sum_\mu w_{\mu i} h_\mu \right] \equiv \Theta \left[g_i + \sum_\mu w_{i\mu} h_\mu \right], \end{aligned} \quad (3)$$

where Θ is the Heaviside function, and Φ_μ is the response function (Fig. 1(b) in main text). Starting from a configuration, such zero-temperature Markov Chain runs until convergence to a local minimum of E .

In practice, to make finite-size corrections to our mean-field theory small, we considered RBMs with up to $N \sim 10^4$ visible units. Such large R-RBM were simulated using a nVidia Tesla K40 GPU, programmed with Theano [7].

D. Finding local maxima of $P[\mathbf{v}]$

Given an RBM with energy defined as above, the marginal $P[\mathbf{v}]$ is characterized by a Gibbs distribution and a free energy :

$$\begin{aligned} P[\mathbf{v}] &= \int \prod_{\mu} dh_{\mu} \frac{1}{Z} e^{-E[\mathbf{v}, \mathbf{h}]} = \frac{1}{Z} e^{-F[\mathbf{v}]} \\ F[\mathbf{v}] &= - \sum_i g_i v_i + \sum_{\mu} U_{\mu}^{eff} \left(\sum_i w_{i\mu} v_i \right) \\ U_{\mu}^{eff}(x) &= -\log \left[\int dh e^{-U_{\mu}(h)+xh} \right] \end{aligned} \quad (4)$$

In order to find the local maxima of $P[\mathbf{v}]$ (*i.e.* the local minima of $F[\mathbf{v}]$) , we modify it by introducing an inverse temperature β :

$$\begin{aligned} P^{\beta}[\mathbf{v}] &= \frac{1}{Z(\beta)} e^{-\beta F[\mathbf{v}]} \\ Z(\beta) &= \sum_{\mathbf{v}} e^{-\beta F[\mathbf{v}]} \end{aligned} \quad (5)$$

Sampling from this distribution at $\beta \neq 1$ is not trivial, as $P^{\beta}[\mathbf{v}]$ is not the marginal distribution of $P^{\beta}[\mathbf{v}, \mathbf{h}]$ when $\beta \neq 1$. While sampling from $P^1[\mathbf{v}]$ is easy, as one can simply sample from the joint distribution $P^1[\mathbf{v}, \mathbf{h}]$ using Gibbs steps, this is not true for $\beta \neq 1$; in particular the local maxima of $P^{\beta}[\mathbf{v}, \mathbf{h}]$ are not equivalent to those of $P^{\beta}[\mathbf{v}]$. We notice however that when $\beta \geq 1$ is an integer, $P^{\beta}[\mathbf{v}]$ can be interpreted as the $\beta = 1$ distribution of another RBM $P'^1[\mathbf{v}]$ with βM hidden units (each hidden unit is replicated β times) and visible fields $g' = \beta g$.

Sampling from such RBM can be done as following :

- Compute the hidden layer inputs $\sum_i w_{i\mu} v_i$
- Sample independently β replicas h_{μ}^r of h_{μ} from $P^1[h_{\mu} | \mathbf{v}]$
- Compute the visible layer inputs $I_i = \sum_{r=1}^{\beta} \sum_{\mu} w_{i\mu} h_{\mu}^r$
- Sample v_i from the Bernoulli distribution $Bern \left[\beta(g_i + \frac{1}{\beta} I_i) \right]$

When $\beta \rightarrow \infty$, $\frac{1}{\beta} \sum_{r=1}^{\beta} h_r$ coincides with the conditional average of \mathbf{h} given \mathbf{v} , $\mathbb{E}[\mathbf{h} | \mathbf{v}]$. Therefore, the zero temperature sampling Gibbs step for the free energy is equivalent to :

$$\begin{aligned} h_{\mu} &\leftarrow \mathbb{E}[h_{\mu} | \mathbf{v}] \\ v_i &\leftarrow \Theta \left[g_i + \sum_{\mu} w_{i\mu} h_{\mu} \right] \end{aligned} \quad (6)$$

III. NUMERICAL PROXIES FOR CONTROL AND ORDER PARAMETERS

Several control and order parameters are well defined for R-RBM in the thermodynamical limit, but not for typical RBM trained on data. For R-RBM instances, the average weight sparsity p is well defined because the weights take only three distinct values $\{-\frac{1}{\sqrt{N}}, 0, \frac{1}{\sqrt{N}}\}$, but for RBM trained on data, the weights $w_{i\mu}$ are never exactly equal to zero. Similarly, the number of strongly activated hidden units L is well-defined for R-RBM in the thermodynamic limit $N \rightarrow \infty$ because their activity scales as \sqrt{N} ; but in general, all hidden units have finite activation. Proxies are therefore required to compare theoretical and numerical results. We consider 'consistent' proxies, giving back (in the large size limit), the original parameters for RBMs drawn from the R-RBM ensemble.

A. Participation Ratios PR

Participation ratios are used to estimate numbers of nonzero components in a vector, while avoiding the use of arbitrary thresholds. The Participation Ratio (PR_a) of a vector $\mathbf{x} = \{x_i\}$ is

$$PR_a(\mathbf{x}) = \frac{(\sum_i |\mathbf{x}_i|^a)^2}{\sum_i |\mathbf{x}_i|^{2a}}$$

If \mathbf{x} has K nonzero and equal (in modulus) components PR is equal to K for any a . In practice we use the values $a = 2$ and 3 : the higher a is, the more small components are discounted against strong components in \mathbf{x} .

B. Number L of active hidden units

In both numerical simulations of R-RBM and on RBM trained on MNIST, we estimate L , for a given hidden-unit configuration \mathbf{h} , through

$$\hat{L} = PR_3(\mathbf{h})$$

To understand the choice $a = 3$, consider a typical activation configuration \mathbf{h} for a R-RBM :

$$h_\mu = \begin{cases} m\sqrt{N} & \text{if } 1 \leq \mu \leq L , \\ \sqrt{r} x_\mu & \text{if } L + 1 \leq \mu \leq M , \end{cases} \quad (7)$$

where the magnetization m and mean square activity r are $\mathcal{O}(1)$, and x_μ are random variables with zero mean, and even moments of the order of unity. The first L hidden units are strongly activated ($\mathcal{O}(\sqrt{N})$ activity), whereas the remaining $N - L$ others are not (activations of the order of 1). Here, we assume L to be finite as $N \rightarrow \infty$. One computes :

$$\begin{aligned} PR_2(h) &\sim \frac{(Lm^2 N + (N - L)r)^2}{Lm^4 N^2 + (N - L)r^2} = L \times \frac{(1 + \frac{N-L}{N} \frac{r}{Lm^2})^2}{1 + \frac{N-L}{N^2} \frac{r^2}{Lm^4}} \xrightarrow[N \rightarrow \infty]{} L(1 + \frac{r}{Lm^2})^2 , \\ PR_3(h) &\sim \frac{(Lm^3 N^{3/2} + (N - L)r^{3/2})^2}{Lm^6 N^3 + (N - L)r^3} = L \times \frac{(1 + \frac{N-L}{N^{3/2}} \frac{r^{3/2}}{Lm^3})^2}{1 + \frac{N-L}{N^3} \frac{r^3}{Lm^6}} \xrightarrow[N \rightarrow \infty]{} L . \end{aligned} \quad (8)$$

Hence choosing coefficient $a = 3$ ensures that the participation ratio (a) does not take into account the weak activations in the thermodynamical limit, and (b) converges to the true value L if all magnetizations are equal.

C. Normalized Magnetizations \tilde{m}

Given a RBM and a visible layer configuration, we define the normalized magnetization of hidden unit μ as the normalized overlap between the configuration and the weights attached to the unit:

$$\tilde{m}_\mu = \frac{\sum_i (2v_i - 1)w_{i\mu}}{\sum_i |w_{i\mu}|} \in [-1, 1]$$

This definition is consistent with the Hopfield model. For R-RBM, we also have, in the thermodynamical limit, $\hat{m}_\mu = \frac{2I_\mu}{p\sqrt{N}}$, where I_μ is the input received by the hidden unit from the visible layer; m_μ is $\mathcal{O}(1)$ for strongly activated hidden units, and $\mathcal{O}(\frac{1}{\sqrt{N}})$ for the others.

For a given configuration \mathbf{v} , with \hat{L} activated hidden units, the normalized magnetization of the activated hidden units $\tilde{m} = \frac{m}{\hat{L}}$ can be estimated as the average of the \hat{L} highest magnetizations \hat{m}_μ .

D. Weight sparsity p

A natural way to estimate the fraction of non-zero weights $w_{i\mu}$ would be to count the number of weights with absolute value above some threshold t . However, there is no simple satisfactory choice for t . Indeed, the fraction of

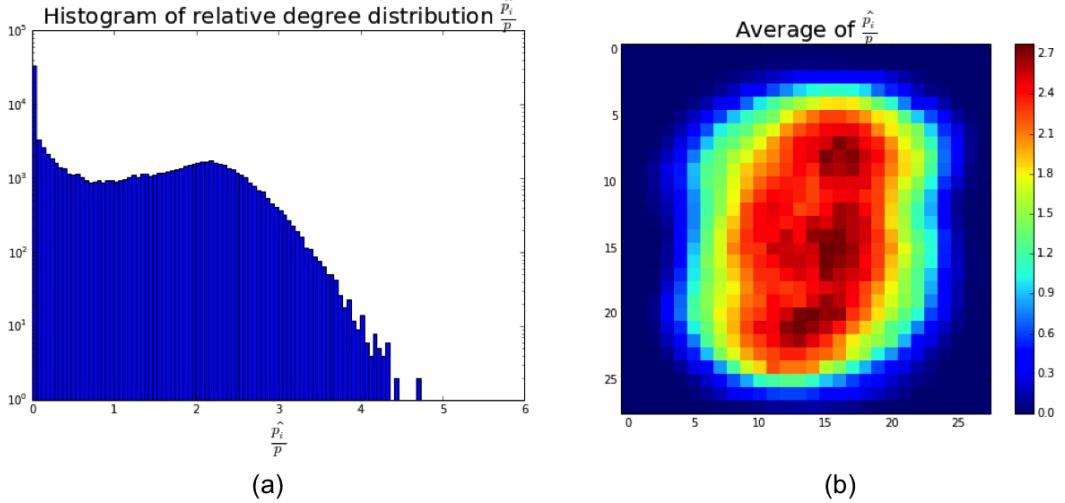


FIG. 5: (a) Histogram of $\tilde{p}_i = \frac{p_i}{p}$ values, across all visible units and RBMs inferred on MNIST. (b) Average across all RBM of $\tilde{p}_i = \frac{p_i}{p}$, for each visible unit

non-zero weights should not depend on the scale of the weights, *i.e.* it should be invariant under the global rescaling transformation $\{w_{i\mu}\} \rightarrow \{\lambda w_{i\mu}\}$. As the scale of weights vary from RBMs to RBMs and, for each RBM, across training it appears difficult to select an appropriate value for t . A possibility would be to use a threshold adapted to each RBM, *e.g.* $t \propto \kappa \sqrt{\frac{W_2}{M}}$, where κ would be some small number. Our experiments show that it is not accurate enough, due to the scale disparities across the hidden-unit weight vectors \mathbf{w}_μ . Rather than adapting thresholds to each hidden unit of each RBM, we use Participation Ratios, which naturally enjoy the scale invariance property. We estimate the fraction of nonzero weights through

$$\hat{p} = \frac{1}{MN} \sum_\mu PR_2(\mathbf{w}_\mu)$$

For R-RBM with $w_{i\mu} \in [-W_0, 0, W_0]$ with corresponding probabilities $[\frac{p}{2}, 1-p, \frac{p}{2}]$, the estimator is consistent: $\hat{p} = p$.

E. Weights heterogeneities

As seen from the features of Fig. 2 in the main text, not all visible units are equally connected to the hidden layer. To better capture this effect, one can study R-RBM with any arbitrary distribution of p_i . Analogously to the homogeneous case a high sparsity limit is obtained when the average sparsity, $p = \frac{1}{N} \sum_i p_i$, vanishes. We define the distribution of the ratios $\tilde{p}_i = \frac{p_i}{p}$ in the $p \rightarrow 0$ limit. In practice the ratios are estimated through

$$\tilde{p}_i = \frac{\sum_i w_{i\mu}^2}{\frac{1}{M} \sum_{i,\mu} w_{i\mu}^2} . \quad (9)$$

For a heterogeneous R-RBM, we have consistently $\tilde{p}_i = \frac{\hat{p}_i}{p} = \frac{p_i}{p}$. Looking at the histogram of values of \tilde{p}_i across all RBM inferred on MNIST, we find a non-negligible spread around one, see Fig. 5. We also display for each visible unit i the average of \tilde{p}_i accross all RBM inferred; we can see that the visible units at the border are indeed the least connected (smaller \tilde{p}_i), whereas the ones at the center are strongly connected (larger \tilde{p}_i).

F. Effective Temperature T

Although RBM distributions are always defined at temperature $T = 1$, the effective temperature is not 1. This is very much like in the Ising model : the behavior of the system depends on an effective temperature $\bar{T} = \frac{T}{J}$ where



FIG. 6: Conditional means $E[\mathbf{v}|\mathbf{h}]$ for two hidden units configurations sampled at equilibrium. Most pixels v_i are frozen, with $E[v_i|\mathbf{h}] \in \{0, 1\}$

J is the coupling strength; a low effective temperature phase correspond to high values of J . For ReLU RBM, the probability distribution of configurations at temperature T is defined as :

$$P_{\mathbf{w}}[\mathbf{v}, \mathbf{h}] = e^{-\frac{E[\mathbf{v}, \mathbf{h}]}{T}} \quad \text{with} \quad \frac{E[\mathbf{v}, \mathbf{h}]}{T} = -\sum_i \frac{g_i}{T} v_i + \sum_\mu \left(\frac{h_\mu^2}{2T} + \frac{h_\mu \theta_\mu}{T} \right) - \sum_{i,\mu} \frac{w_{i\mu}}{T} v_i h_\mu . \quad (10)$$

Let $\bar{\mathbf{h}} = \frac{\mathbf{h}}{\sqrt{T}}$. The probability can be rewritten as $P[\mathbf{v}, \bar{\mathbf{h}}] = e^{-\bar{E}[\mathbf{v}, \bar{\mathbf{h}}]}$ with

$$\bar{E}(\mathbf{v}, \bar{\mathbf{h}}) = -\sum_i \frac{g_i}{T} v_i + \sum_\mu \left(\frac{\bar{h}_\mu^2}{2} + \bar{h}_\mu \frac{\theta_\mu}{\sqrt{T}} \right) - \sum_{i,\mu} \frac{w_{i\mu}}{\sqrt{T}} v_i \bar{h}_\mu . \quad (11)$$

Since the marginal $P[\mathbf{v}]$ is not affected by the change of variable, a ReLU RBM at temperature T is therefore equivalent to another ReLU RBM at temperature $T = 1$, with new fields, thresholds and weights : $\bar{\mathbf{g}} = \frac{\mathbf{g}}{T}$, $\bar{\theta} = \frac{\theta}{\sqrt{T}}$, $\bar{\mathbf{w}} = \frac{\mathbf{w}}{\sqrt{T}}$. Therefore, changing the temperature is equivalent to rescaling the parameters, and in turn, the effective temperature of a given RBM can be deduced from the amplitude of its weights. For a R-RBM at temperature T :

$$W_2 = \frac{1}{M} \sum_{\mu,i} \bar{w}_{i\mu}^2 \underset{N \rightarrow \infty}{\sim} \frac{p}{T} .$$

We therefore estimate the temperature of a given RBM through

$$\hat{T} = \frac{\hat{p}}{\frac{1}{M} \sum_{\mu,i} w_{i\mu}^2} .$$

From this definition, it can be seen that the low temperature regime of the compositional regime, $T \ll p$, is equivalent to $W_2 \gg 1$. In RBM trained on MNIST, we typically find $W_2 \sim 7$

G. Fields g

Similarly to the weights, the fields g_i and normalized fields could be estimated respectively as:

$$\begin{aligned} \hat{g}_i &= \hat{T} \bar{g}_i \\ \hat{\tilde{g}}_i &= \frac{\hat{T}}{\hat{p}} \bar{g}_i = \frac{\bar{g}_i}{\frac{1}{M} \sum_{\mu,i} w_{i\mu}^2} \end{aligned} \quad (12)$$

A naive estimate for the normalized field \tilde{g} would be to average the fields: $\hat{\tilde{g}} = \frac{1}{N} \sum_i \hat{g}_i$. It is however not really meaningful, as the \hat{g}_i are extremely heterogeneous: for instance, the mean value over the sites i of a single RBM is

equal to -0.48 , and is comparable to the standard deviation, 0.40 . This range of variation spans all the phases of R-RBM. To achieve quantitative predictions, we instead adjust the R-RBM parameter g so that q , the mean value of v_i in the visible layer, averaged over thermal fluctuations and quenched disorder, matches the value 0.132 obtained from MNIST data. For the plots of Figure 4 in the main text, this gives $\frac{g}{\hat{p}} = -0.1725$ for homogeneous R-RBM, and $\frac{g}{\hat{p}} = -0.21$ for heterogeneous R-RBM.

H. Thresholds θ

The thresholds and normalized thresholds can be estimated as

$$\begin{aligned}\hat{\theta}_\mu &= \sqrt{\hat{T}} \bar{\theta}_\mu \\ \hat{\tilde{\theta}}_\mu &= \sqrt{\frac{\hat{T}}{\hat{p}}} \bar{\theta}_\mu = \frac{\bar{\theta}_\mu}{\sqrt{\frac{1}{M} \sum_{\mu,i} w_{i\mu}^2}}\end{aligned}\quad (13)$$

Again, a naive estimate for the normalized threshold $\tilde{\theta}$ would be the average $\hat{\tilde{\theta}} = \frac{1}{M} \sum_\mu \hat{\theta}_\mu$ but this estimate is not meaningful. Indeed, contrary to the R-RBM case, the inputs I_μ of the hidden units μ are not evenly distributed around zero: $\mathbb{E}[I_\mu] \neq 0$. Hence, even if the threshold is equal to zero, the activation probability can be different from 0.5 . We take this effect into account by subtracting the average value of the inputs from the average of θ , and find that the difference is equal to 0.33 , with standard deviation 1.11 . This range of value for θ again spans all phases. In order to use a well-defined value, we choose θ such that the critical capacity $a_c^{R-RBM}(\ell_{max}) = 0.5$, where $\ell_{max} \sim 1.5$ is the maximum average index number observed across all RBMs trained for Fig. 4 in the main text. This estimation gives $\hat{\tilde{\theta}} \sim 1.5$.

- [1] A. Fischer, C. Igel. *Iberoamerican Congress on Pattern Recognition*, pp. 14-36 (2012).
- [2] G. Hinton. *Momentum* **9**, 926 (2010).
- [3] Tieleman, T. *Proceedings of the 25th international conference on Machine learning*, pp. 1064-1071 (2008, July).
- [4] Desjardins, G., Courville, A., Bengio, Y., Vincent, P., Delalleau, O. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 145-152 (2010).
- [5] J. Tubiana, R. Monasson, *in preparation* (2017).
- [6] Salakhutdinov, R., & Murray, I. *In Proceedings of the 25th international conference on Machine learning* (pp. 872-879). (2008, July)
- [7] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions, *arXiv preprint arXiv:1605.02688*. (2016).