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 :

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

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 $\lambda_i = 5\ 10^{-3}$, decaying geometrically (decay starts after 60 epochs) to $\lambda_f = 5\ 10^{-4}$

PCD is known to be inaccurate toward the end of learning, because the parameters evolve too fast with respect to the 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

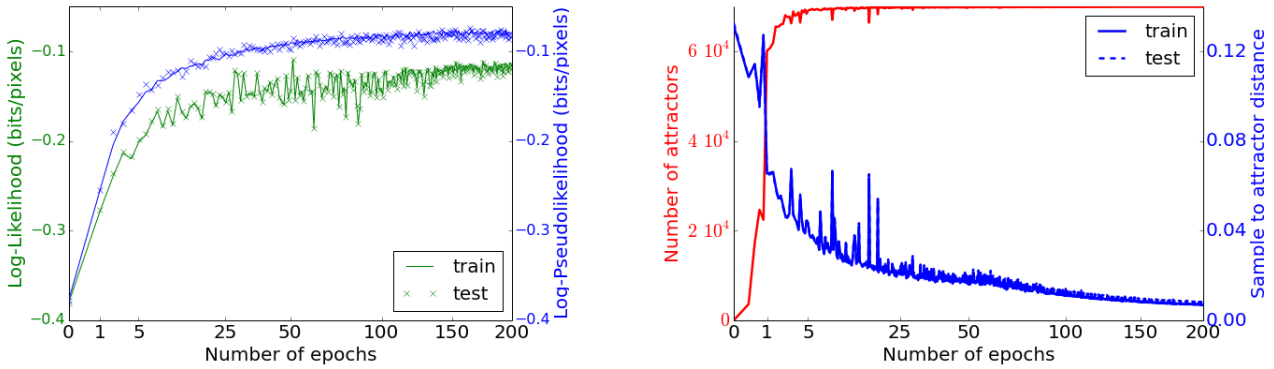

FIG. 1: **(a)** Evolution throughout training of the data loglikelihood (left scale) and pseudo-loglikelihood (right scale) computed over the training and test sets. **(b)** Evolution of the number of distinct local maxima of $P_W(v)$ (left scale) and the distance to the original sample (right scale, 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. Controling weight sparsity with regularization

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

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

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.

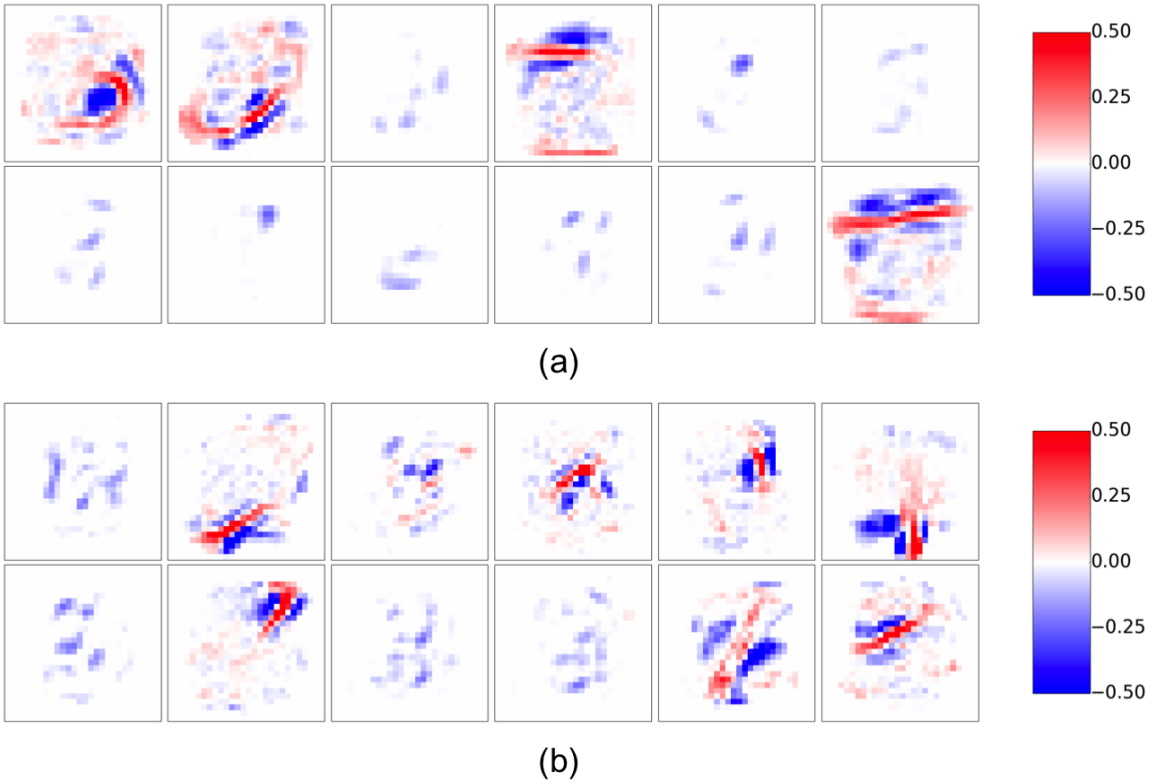

FIG. 2: Subset of 12 weight features produced by training on MNIST, regularized with $L^1, \lambda_1 = 10^{-3}$ (top panel), and $L^2, \lambda_2 = 3.10^{-5}$ (bottom panel). Both have overall sparsity $p \sim 0.036$, but the latter is more homogeneously distributed across hidden units.

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.

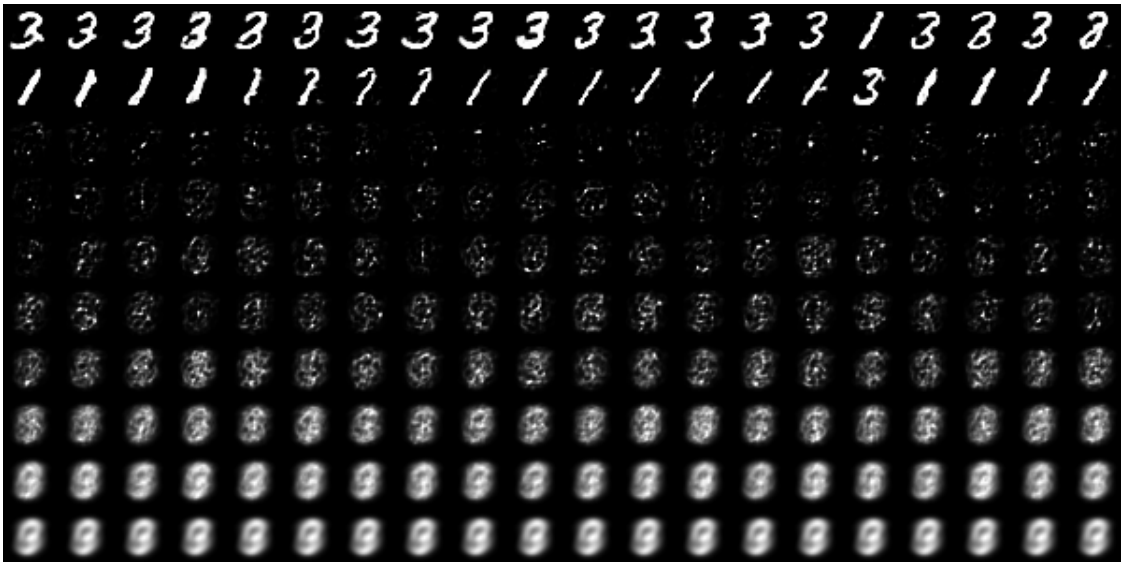

FIG. 4: Ten Monte Carlo Markov Chains realizations at different inverse temperatures, coupled by replica exchange. The plots shows the conditional expectations of visibile units, $\mathbb{E}[\mathbf{v}|\mathbf{h}]$, for thermalized hidden-unit activities, $\mathbf{h}$.

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 $\sim$ 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 $\beta \to \infty$, a dynamical Gibbs step becomes deterministic :

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

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 :

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

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:

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

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 interpretated 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 $\beta \to \infty$, $\frac{1}{\beta}\sum_{r=1}^{\beta} h_r$ 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 :

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

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 $N \to \infty$ 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

$$PR_a(\mathbf{x}) = \frac{(\sum_{\mathbf{i}} |\mathbf{x_i}|^{\mathbf{a}})^{\mathbf{2}}}{\sum_{\mathbf{i}} |\mathbf{x_i}|^{\mathbf{2a}}}$$

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 :

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

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 $N \to \infty$. One computes :

$$\begin{aligned} PR_2(h) &\sim \frac{(Lm^2N + (N-L)r)^2}{Lm^4N^2 + (N-L)r^2} = L \times \frac{(1 + \frac{N-L}{N}\frac{r}{Lm^2})^2}{1 + \frac{N-L}{N^2}\frac{r^2}{Lm^4}} \underset{N\to\infty}{\longrightarrow} L(1 + \frac{r}{Lm^2})^2 \ , \\ PR_3(h) &\sim \frac{(Lm^3N^{3/2} + (N-L)r^{3/2})^2}{Lm^6N^3 + (N-L)r^3} = L \times \frac{(1 + \frac{N-L}{N^{3/2}}\frac{r^{3/2}}{Lm^3})^2}{1 + \frac{N-L}{N^3}\frac{r^3}{Lm^6}} \underset{N\to\infty}{\longrightarrow} L \ . \end{aligned} \tag{8}$$

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 $\tilde{m} = \frac{m}{p/2}$ can be estimated as the average of the $\hat{L}$ highest magnetizations $\hat{m}_\mu$.

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

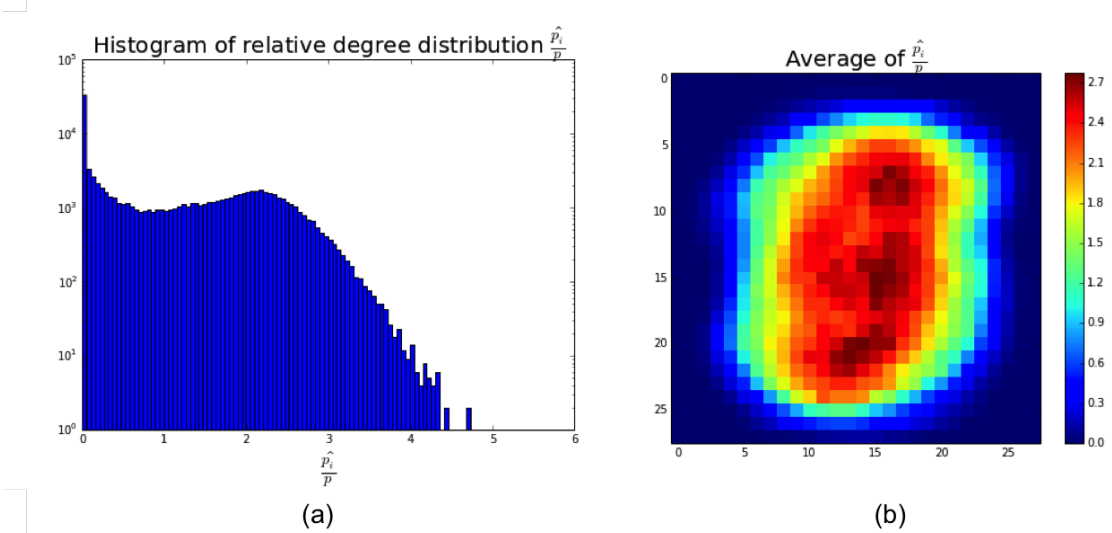

FIG. 5: (a) Histogram of $\tilde{p}_i = \frac{p_i}{p}$ values, across all visible units and RBMs inferred on MNIST. (b) Average across all RBM of $\tilde{p}_i = \frac{p_i}{p}$, for each visible unit

non-zero weights should not depend on the scale of the weights, *i.e.* it should be invariant under the global rescaling transformation $\{w_{i\mu}\} \to \{\lambda\, w_{i\mu}\}$. 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 $\mathbf{w}_\mu$. 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 $p \to 0$ limit. In practice the ratios are estimated through

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

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 $\hat{T} = \frac{T}{J}$ where

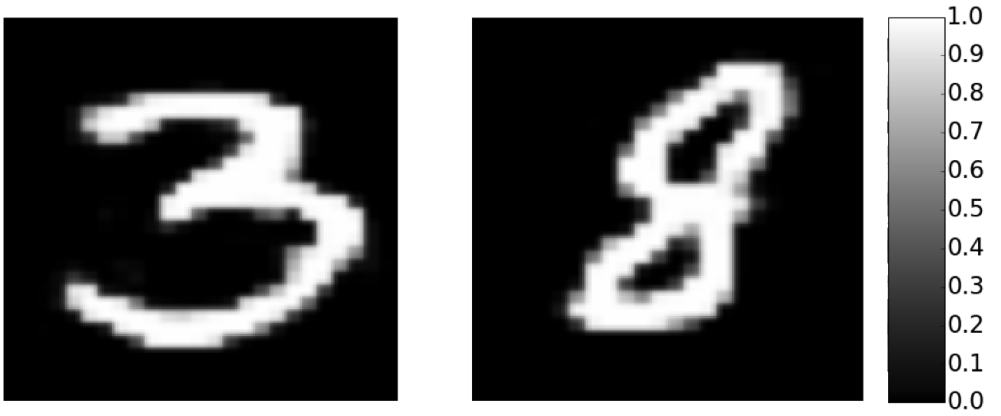

FIG. 6: Conditional means $\mathbb{E}[\mathbf{v}|\mathbf{h}]$ for two hidden units configurations sampled at equilibrium. Most pixels v_i are frozen, with $\mathbb{E}[v_i|\mathbf{h}] \in \{0, 1\}$

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

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

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

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

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:

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

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:

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

A naive estimate for the normalized field $\tilde{g}$ would be to average the fields: $\hat{\tilde{g}} = \frac{1}{N}\sum_i \hat{\tilde{g}}_i$. It is however not really meaningful, as the $\hat{\tilde{g}}_i$ 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 $\frac{\hat{g}}{\hat{p}} = -0.1725$ for homogeneous R-RBM, and $\frac{\hat{g}}{\hat{p}} = -0.21$ for heterogeneous R-RBM.

H. Thresholds θ

The thresholds and normalized thresholds can be estimated as

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

Again, a naive estimate for the normalized threshold $\tilde{\theta}$ would be the average $\hat{\tilde{\theta}} = \frac{1}{M}\sum_\mu \hat{\tilde{\theta}}_\mu$ 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 substracting 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 $\alpha_c^{R-RBM}(\ell_{max}) = 0.5$, 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).