Emergence of Compositional Representations in Restricted Boltzmann Machines

J. Tubiana[*] and R. Monasson

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

(Received 21 November 2016; published 28 March 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 data set MNIST.

DOI: 10.1103/PhysRevLett.118.138301

Recent years have witnessed major progress in supervised machine learning, e.g., in video, audio, and 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 complex data manifolds well in high-dimensional spaces [2]. Once learned, such generative models can be used for denoising, completion, artificial data generation, etc. 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) 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; owing to the bidirectionality of the 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 $i = 1, \ldots, N$, 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 $\mathcal{U}_\mu$ is a potential acting on hidden unit μ; due 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 with 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

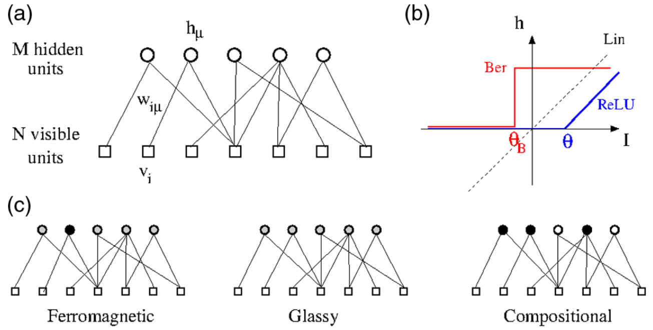

FIG. 1. (a) Architecture of RBM. Visible (v_i, $i = 1, \ldots, N$) and hidden (h_μ, $\mu = 1, .., M$) units are connected through weights ($w_{i\mu}$). (b) Activation functions Φ of Bernoulli, linear, and rectified linear units. The corresponding potentials are $\mathcal{U}^{\rm Lin}(h) = (h^2/2)$; $\mathcal{U}^{\rm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

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 $\mathbf{w}_\mu$ 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 data set 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 [24], Sec. I. Figure 2(a) shows typical features $\mathbf{w}_\mu = \{w_{i\mu}\}$ 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 the 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 [24], Fig. 1(b)], hence showing the generative power of RBM.

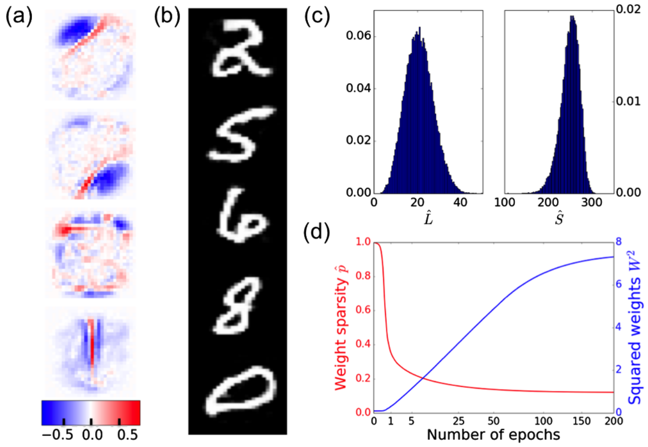

FIG. 2. Training of RBM on MNIST, with $N = 28 \times 28$ visible units and $M = 400$ hidden ReLU. (a) Set of weights $\mathbf{w}_\mu$ attached to four hidden units μ. (b) Averages of $\mathbf{v}$ conditioned to five hidden-unit configurations $\mathbf{h}$ sampled from the RBM at equilibrium. Black and 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.

Learning is accompanied by structural changes in RBM, which we track with two parameters: $\hat{p} = (1/MN) \times \sum_\mu [(\sum_i w_{i\mu}^2)^2 / \sum_i w_{i\mu}^4]$ and $W_2 = (1/M) \sum_{i,\mu} w_{i\mu}^2$. Those parameters are proxies for, respectively, the fraction of nonzero weights and the effective inverse temperature; see Supplemental Material [24], Sec. 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 with large energy costs. Visible units are effectively at very low temperature, as can be seen from the quasibinary nature of conditional averages in Fig. 2(b) and Supplemental Material [24], 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 combinations 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 favored 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 $-(1/\sqrt{N})$, 0, $+(1/\sqrt{N})$ with probabilities equal to, respectively, $(p_i/2)$, $1 - p_i$, $(p_i/2)$; $p_i \in [0; 1]$ 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 $\mathcal{U}_\mu = \mathcal{U}^{\mathrm{ReLU}}$ [Fig. 1(a)]. We define the ratio of the numbers of hidden and visible units, $\alpha = M/N$.

Given a visible layer configuration $\mathbf{v}$, hidden units μ coding for features $\mathbf{w}_\mu$ present in $\mathbf{v}$ are strongly activated: their inputs $I_\mu = \mathbf{w}_\mu \cdot \mathbf{v}$ are strong and positive, comparable to the product of the norms of $\mathbf{w}_\mu$ ($\simeq \sqrt{p}$ for large N) and $\mathbf{v}$ (of the order of $\sqrt{pN}$), and, hence, 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)] coincides with the number L of strongly activated units when $N \to \infty$; choosing exponent $a = 2$ in $\hat{L}$ rather than $a = 3$ would have introduced biases coming from weakly activated units (Supplemental Material [24], Sec. 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

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

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, $q = \sum_i p_i \langle v_i \rangle_{\mathrm{GS}}/(Np)$ 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 [19]. In (2) $Dz = (dz/\sqrt{2\pi})e^{-z^2/2}$ 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 nonmagnetized 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}.

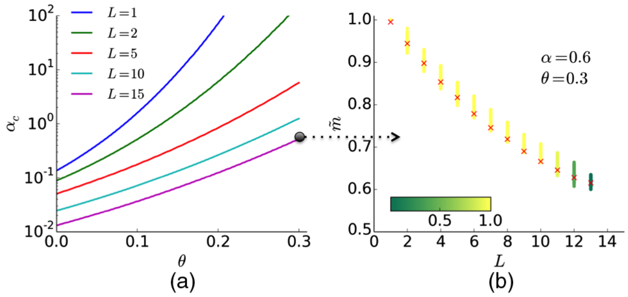

FIG. 3. Compositional regime in R-RBM. (a) Critical lines in the θ, α plane below which L hidden units may be strongly activated, calculated from the optimization of E_{GS} (2). Parameters are $p_i = p = 0.1$, $g = -0.02$. (b) Comparison of theoretical (red crosses) and numerical simulations ($N = 10^4$ 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 $L = 1, 2, \ldots$ features, were launched; the color code indicates the probability that the same L hidden units are magnetized after convergence (see the 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 behavior, which we call hereafter the 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 $e_\ell = E_{\mathrm{GS}}(L = \ell/p)/p$ 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, $m = (p/2)\tilde{m}$, vanish linearly with p [20]. Nonmagnetized hidden units have activities of the order of $\sqrt{r} \sim \sqrt{p}$, and can be shut down 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

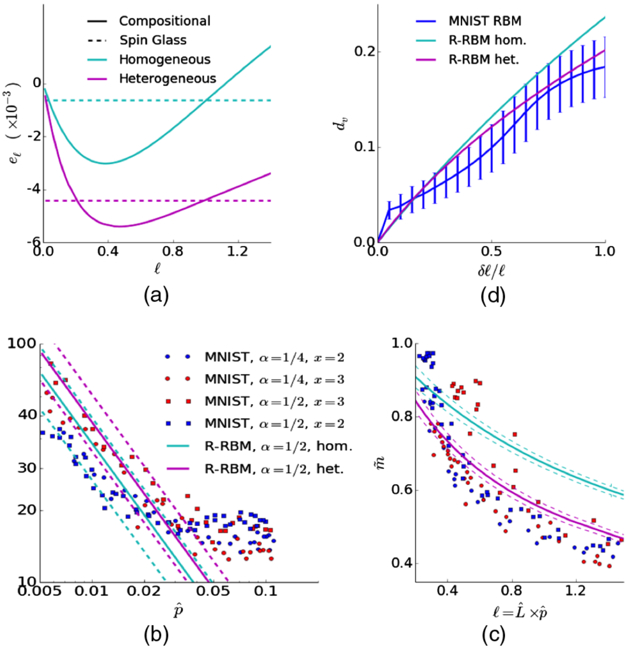

FIG. 4. (a) Behavior of the GS energy e_ℓ vs $\ell = L \times p$ in the $p \to 0$ limit. Parameters: $\tilde{\theta} = 1.5$, $\alpha = 0.5$. [(b)–(d)] Quantitative predictions in the compositional regime of R-RBM compared to RBMs inferred on MNIST. Each point represents a ReLU RBM trained with various 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 $\tilde{m}$ 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 seven pixels, both on the training and test sets; see Supplemental Material [24], Fig. 1. In all four panels predictions for the homogeneous ($\tilde{g} = -0.21$) and heterogeneous ($\tilde{g} = -0.1725$; see Supplemental Material [24], Sec. III E) cases are shown in, respectively, cyan and magenta.

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 monotonically 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 (α, $\tilde{g} = (g/p)$, $\tilde{\theta} = (\theta/\sqrt{p})$) will be detailed in [19]. 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.

Characteristic properties of our compositional phase are confronted to ReLU RBMs trained on MNIST in Figs. [4(b) and 4(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 training 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, $\tilde{m}$. Figure 4(b) shows $\hat{L}$ vs $\hat{p}$, in good agreement with the R-RBM theoretical scaling $L \sim (\ell^\star/p)$. Figure 4(c) shows that $\tilde{m}$ 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 a 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 [24], Sec. III E). Its GS energy can be calculated from (2); see Fig. 4(a) [19]. Results are shown in Figs. 4(b) and 4(c) to be in good agreement with RBM trained on MNIST.

RBMs, 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 [19], 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 monotonically 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. Markov Chain Monte Carlo (MCMC) of MNIST-trained RBM (videos available in Supplemental Material [24]) 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 the energy much, 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). A finite capacity at zero temperature could be achieved only in the limit case of extreme sparsity $p = c/N$, giving $\alpha \sim c^{-2}$ [22]; for typical MNIST values $p \simeq 0.1$ and $N = 784$ this would give $\alpha \sim 2 \times 10^{-4}$. 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 [17,18,22]) 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 layer 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.

We are grateful to C. Fisher and G. Semerjian for useful discussions. This work was partly funded by the Centre National de la Recherche Scientifique (CNRS)-Inphyniti Inferneuro and the Human Frontier Science Program (HFSP) Grant No. RGP0057/2016, and benefited from the support of NVIDIA Corporation with the donation of a Tesla K40 GPU card.

*jerome.tubiana@ens.fr

[1] Y. LeCun, Y. Bengio, and J. Hinton, Nature (London) **521**, 436 (2015).

[2] Y. Bengio, A. Courville, and P. Vincent, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[3] P. Smolensky, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (MIT Press, 1986), Chap. 6, pp. 194–281.

[4] R. Salakhutdinov, A. Mnih, and G. Hinton, in *Proceedings of the 24th International Conference on Machine Learning* (ACM, Corvalis, Oregon, 2007), pp. 791–798.

[5] G. Hinton, Momentum **9**, 926 (2010).

[6] T. Tieleman, in *Proceedings of the 25th International Conference on Machine Learning* (ACM, Helsinki, Finland, 2008), pp. 1064–1071.

[7] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, and O. Delalleau, in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (ACM, Pittsburgh, PA, 2010), pp. 145–152.

[8] M. Gabrie, E. W. Tramel, and F. Krzakala, *Advances in Neural Information Processing Systems* (The MIT Press, Montreal, Quebec, 2015), pp. 640–648.

[9] A. Fischer and C. Igel, in *Iberoamerican Congress on Pattern Recognition* (Springer, Buenos Aires, Argentina, 2012), pp. 14–36.

[10] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, Neural Netw. **34**, 1 (2012).

[11] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

[12] V. Nair and G. E. Hinton, in *Proceedings of the 27th International Conference on Machine Learning* (Omnipress, Haïfa, Israël, 2010), pp. 807–814.

[13] A. Treves, J. Comput. Neurosci. **2**, 259 (1995).

[14] Y. LeCun, C. Cortes, and C. J. Burges, *The MNIST Database of Handwritten Digits* (1998).

[15] B. A. Olshausen and D. J. Field, Nature (London) **381**, 607 (1996).

[16] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. Lett. **55**, 1530 (1985).

[17] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, and F. Moauro, Phys. Rev. Lett. **109**, 268101 (2012).

[18] E. Agliari, A. Annibale, A. Barra, A. C. C. Coolen, and D. Tantari, J. Phys. A **46**, 415003 (2013).

[19] J. Tubiana and R. Monasson (to be published).

[20] 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 [19].

[21] M. Tsodyks and M. V. Feigel'man, Europhys. Lett. **6**, 101 (1988).

[22] P. Sollich, D. Tantari, A. Annibale, and A. Barra, Phys. Rev. Lett. **113**, 238106 (2014).

[23] D. Krotov and J. J. Hopfield, arXiv:1606.01164 [NIPS (to be published)].

[24] See Supplemental Material http://link.aps.org/supplemental/10.1103/PhysRevLett.118.138301 for details on training RBMs, sampling from RBMs, and estimating control and order parameters. Videos of Monte Carlo simulations are also provided, which includes Refs. [25,26].

[25] R. Salakhutdinov and I. Murray, in *Proceedings of the 25th International Conference on Machine Learning* (2008), pp. 872–879.

[26] Theano Development Team, arXiv:1605.02688.