# Exact Training of Restricted Boltzmann Machines on Intrinsically Low Dimensional Data

A. Decelle[1,2] and C. Furtlehner[3,1,*]

[1]*LISN, AO team, Bât 660 Université Paris-Saclay, Orsay Cedex 91405, France*
[2]*Departamento de Física Téorica I, Universidad Complutense, 28040 Madrid, Spain*
[3]*Inria Saclay-Tau team, Bât 660 Université Paris-Saclay, Orsay Cedex 91405, France*

(Received 19 March 2021; revised 22 July 2021; accepted 14 September 2021; published 7 October 2021)

The restricted Boltzmann machine is a basic machine learning tool able, in principle, to model the distribution of some arbitrary dataset. Its standard training procedure appears, however, delicate and obscure in many respects. We bring some new insights to it by considering the situation where the data have low intrinsic dimension, offering the possibility of an exact treatment and revealing a fundamental failure of the standard training procedure. The reasons for this failure—like the occurrence of first-order phase transitions during training—are clarified thanks to a Coulomb interactions reformulation of the model. In addition, a convex relaxation of the original optimization problem is formulated, thereby resulting in a unique solution, obtained in precise numerical form on $d = 1$, 2 study cases, while a constrained linear regression solution can be conjectured on the basis of an information theory argument.

DOI: 10.1103/PhysRevLett.127.158303

Recent advances in machine learning (ML) pervade now many other scientific domains including physics by providing new powerful data analysis tools in addition to traditional statistical ones. The restricted Boltzmann machine (RBM) could be considered as one of these when already a large spectrum of possible uses has been proposed in physics [1–5]. Introduced more than three decades ago [6], the RBM played an important role in early developments of deep learning [7]. It is a special case of generative models [8–10] that remains very popular thanks to its simplicity and effectiveness when applied to moderately high dimensional data [11–13]. It is a two-layers undirected neural network that represents the data in the form of a Gibbs distribution of visible and latent variables (see Fig. 1),

$$p(\mathbf{s},\boldsymbol{\sigma}) = \frac{1}{Z[\Theta]}\exp\left(\sum_{i,j} s_i W_{ij}\sigma_j - \sum_{i=1}^{N_v}\eta_i s_i - \sum_{j=1}^{N_h}\theta_j\sigma_j\right). \tag{1}$$

The former noted $\mathbf{s} = \{s_i, i = 1, \ldots, N_v\}$ correspond to explicit representations of the data, while the latter noted $\boldsymbol{\sigma} = \{\sigma_j, j = 1, \ldots, N_h\}$ are there to build arbitrary dependencies among the visible units. They play the role of an interacting field among visible nodes. While many different types of variables can be considered, we take here spin variables $s_i, \sigma_j \in \{-1, 1\}$ for definiteness. $\Theta = (W, \boldsymbol{\eta}, \boldsymbol{\theta})$ are the parameters, $W$ being the weight matrix, $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are local field vectors called, respectively, visible and hidden biases. Each weight vector associated with a given hidden unit and its corresponding bias defines an hyperplane partitioning the visible space into two regions corresponding to the hidden unit being activated or not (see Fig. 1). $Z[\Theta]$ is the partition function of the system. The joint distribution between visible variables is then obtained by summing over hidden ones. Learning the RBM amounts to find $\Theta$ such that generated data obtained by sampling this distribution should be statistically similar to the training data. The standard method to infer the parameters is to maximize the log-likelihood (LL) of the model

$$\begin{aligned}\mathcal{L}[\Theta] = \sum_j \left\langle \log\cosh\left(\sum_i W_{ij}s_i - \theta_j\right)\right\rangle_{\text{Data}} \\ - \sum_i \eta_i \langle s_i\rangle_{\text{Data}} - \log(Z[\Theta]),\end{aligned} \tag{2}$$

with $\langle\rangle_{\text{Data}}$ denoting the average over training data. This is a nontrivial optimization problem in two respects: it is

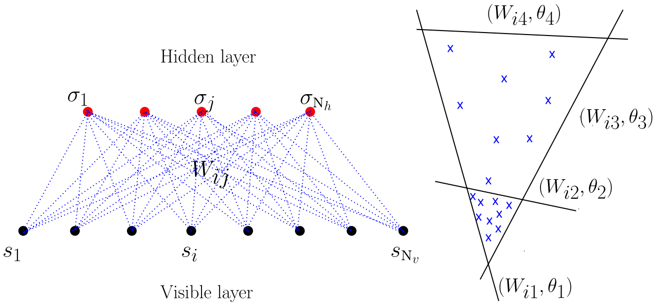

FIG. 1. Bipartite structure of the RBM (left). Hyperplanes defined by the weight vectors and bias associated with each hidden variable can delimit fixed density regions in input space (right).

nonconvex and the loss function $-\mathcal{L}[\Theta]$ is difficult to estimate because $\log(Z[\Theta])$ is not tractable. Nevertheless, the gradient $\nabla_\Theta \mathcal{L}[\Theta]$ can be written in terms of simple response functions of the RBM. These can be estimated approximately via Monte Carlo methods, leading to various algorithms called contrastive divergence [14] with possible refinements [15,16].

The similarity of the RBM with disordered spin systems has raised a lot of interest in statistical physics. Mean-field-based training algorithms and analyses have been proposed [17–20], a mapping with the Hopfield model has been found in [21], retrieval capacity has been characterized in [22,23], and compositional mechanisms are analyzed in [24,25] (see more recent references, e.g., in [26]).

In previous works [27,28] we studied to what extent the learning process of the RBM is reflected in the spectral dynamics of the weight matrix: a certain number of modes, corresponding to principal modes of the data, emerge from a Marchenko-Pastur bulk at initialization and condense to build up a structured ferromagnetic phase. Here we focus on the latter and most difficult stage and show that the two main difficulties (nontractability and nonconvexity) of the training can be addressed in the special case, where a flat intrinsic space of low dimension has been identified in the first stage.

*Effective theory in the ferromagnetic phase.*—Let us first disentangle the contribution of the collective modes corresponding to the information stored from the data (the ferromagnetic and difficult part) from the other degrees of freedom corresponding to the noise (the paramagnetic and easy part). After summing over the hidden variables in (1), the visible distribution reads

$$P[\mathbf{s}|\Theta] = \frac{1}{Z[\Theta]} \exp\left[\sum_{j=1}^{N_h} \log\cosh\left(\sum_{i=1}^{N_v} W_{ij} s_i - \theta_j\right) - \sum_i \eta_i s_i\right]. \tag{3}$$

As in [28] the weight matrix is expressed via its singular value decomposition (SVD)

$$W_{ij} = \sum_{\alpha=1}^{\min(N_v,N_h)} w_\alpha u_i^\alpha v_j^\alpha,$$

with $w_\alpha$, $\boldsymbol{u}^\alpha$, and $\boldsymbol{v}^\alpha$ representing, respectively, the singular values and left and right singular vectors. Assume that some modes $\alpha \in \{1, \ldots, d\}$ have condensed along a magnetization vector denoted $\mathbf{m} = (m_1, \ldots, m_d)$, i.e., that $s_\alpha = m_\alpha = \mathcal{O}(1)$ with $s_\alpha$ defined as

$$s_\alpha \overset{\text{def}}{=} \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} s_i u_i^\alpha.$$

For a RBM trained on some data, $d$ would represent their intrinsic dimension at least locally. The corresponding modes $u_i^\alpha$ can, in principle, be obtained directly from the SVD of the data or emerge naturally from the linear regime of the learning process described in [28]. These magnetization constraints define a canonical statistical ensemble. We look for a change of variables $\mathbf{s} \to (\mathbf{m}, \mathbf{s}^\perp)$, where the original spin variables are replaced by a set of $d$ continuous variables and $\mathcal{N}[\mathbf{m}]$ transverse weakly interacting spin variables. $\mathcal{N}[\mathbf{m}]$ is related to the configurational entropy per spin $\mathcal{S}[\mathbf{m}] = (\mathcal{N}[\mathbf{m}]/N_v)\log(2)$ under these constraints. Thanks to a large deviation argument, $\mathcal{S}[\mathbf{m}]$ is the Legendre transform of (see Supplemental Material [29])

$$\Phi[\boldsymbol{\mu}] = \frac{1}{N_v} \sum_i \log\cosh\left(\sqrt{N_v} \sum_{\alpha=1}^{d} u_i^\alpha \mu_\alpha\right),$$

with $\boldsymbol{\mu}[\mathbf{m}]$ given implicitly by the constraints [30]

$$m_\alpha = \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} u_i^\alpha \tanh\left(\sqrt{N_v} \sum_{\beta=1}^{d} u_i^\beta \mu_\beta\right), \qquad \alpha = 1, \ldots, d. \tag{4}$$

Given a condensed magnetization vector $\mathbf{m}$, there remains $\mathcal{N}[\mathbf{m}]$ interacting degrees of freedom represented by spin variables denoted $\{s_1^\perp, \ldots, s_{\mathcal{N}[\mathbf{m}]}^\perp\}$. With help of this new set of visible variables, the partition function takes the form of a $d$-dimensional integral

$$Z[\Theta] = \int_{\mathcal{D}\subset[-1,1]^d} d^d\mathbf{m}\, e^{-N_v \mathcal{F}[\mathbf{m}|\Theta]}, \tag{5}$$

where the canonical free energy $\mathcal{F}[\mathbf{m}|\Theta] = \mathcal{F}^\parallel[\mathbf{m}|\Theta] + \mathcal{F}^\perp[\mathbf{m}|\Theta]$ is decomposed into two contributions coming, respectively, from the condensed modes and the transverse fluctuations (see Supplemental Material [29])

$$\mathcal{F}^\parallel[\mathbf{m}|\Theta] = -\mathcal{S}[\mathbf{m}] - \sum_{\alpha=1}^{d} \eta_\alpha m_\alpha - V[\mathbf{m}|\Theta], \tag{6}$$

$$\mathcal{F}^\perp[\mathbf{m}|\Theta] = -\frac{1}{N_v} \log\left(\frac{1}{2^{\mathcal{N}[\mathbf{m}]}} \sum_{\mathbf{s}^\perp} e^{-\mathcal{H}_{\text{eff}}[\mathbf{s}^\perp|\mathbf{m},\Theta]}\right), \tag{7}$$

($\eta_\alpha \overset{\text{def}}{=} \frac{1}{\sqrt{N_v}} \sum_i \eta_i u_i^\alpha$), which are, respectively, associated with a potential function for the magnetizations

$$V[\mathbf{m}|\Theta] = \frac{1}{N_v} \sum_{j=1}^{N_h} \log\cosh\left(\sqrt{N_v} \sum_{\alpha=1}^{d} w_\alpha m_\alpha v_j^\alpha - \theta_j\right), \tag{8}$$

and an effective Hamiltonian $\mathcal{H}_{\text{eff}}$ for the transverse degrees of freedom given in the form of a disordered Ising model of $\mathcal{N}[\mathbf{m}]$ spins with paramagneticlike state of order defined for each $\mathbf{m}$ (see Supplemental Material [29]). The default entropy $[\mathcal{N}[\mathbf{m}]\log(2)]$ of the transverse variables is assigned by convenience to $\mathcal{F}^{\parallel}$ so that $\mathcal{F}^{\perp}$ vanishes when $\mathcal{H}_{\mathrm{eff}} = 0$. In the following, we focus on the dominant aspects of the training process resulting from the expression $\mathcal{F}^{\parallel}$. We leave aside specific training problems associated with the transverse fluctuations, like, e.g., the emergence of spurious modes, which will be analyzed elsewhere in detail thanks to this effective Hamiltonian formalism.

*Coulomb formulation and linear regression.*—The potential term in $\mathcal{F}^{\parallel}$, which acts on the magnetization $\mathbf{m}$ representing here the position of a particle in a $d$-dimensional space, can be rewritten as (see Supplemental Material [29])

$$V[\mathbf{m}|\Theta] = \int d\mathbf{n} dz q(\mathbf{n}, z)|\mathbf{n}^T\mathbf{m} - z|, \qquad (9)$$

after introducing in the space $O(d) \times \mathbb{R}$, the density

$$q(\mathbf{n}, z) = \frac{2}{N_v}\sum_{j=1}^{N_h} \nu_j \delta_{\nu_j}\left(z - \frac{\theta_j}{\nu_j}\right)\delta(\mathbf{n} - \mathbf{n}_j) \geq 0, \qquad (10)$$

of latent features, $\delta_\nu(x) = (\nu/2)[1 - \tanh^2(\nu x)]$ being a "smoothed" delta function of width $\nu^{-1}$, with

$$\nu_j = \sqrt{N_v \sum_{\alpha=1}^{d} w_\alpha^2 v_j^{\alpha 2}}, \qquad (11)$$

$$n_j^\alpha = \frac{\sqrt{N_v}}{\nu_j} w_\alpha v_j^\alpha. \qquad (12)$$

The kernel $|\mathbf{n}_j^T\mathbf{m} - z|$ represents the Coulomb potential exerted by a uniformly charged hyperplane, defined by its normal vector $\mathbf{n}$ and its distance $z$ to the origin, on a charge located at $\mathbf{m}$. As a result, each feature $j$ corresponds also to a charged hyperplane of normal vector $\mathbf{n}_j$, offset $z_j = \theta_j/\nu_j$, and finite thickness $\nu_j^{-1}$. At this point, let us remark that the $w_\alpha$ control through (11) both the strength of the Coulomb interaction via (9) and (10) and the charged hyperplanes thickness; the right singular vectors projections $v_j^\alpha$ control on their side the orientation of these hyperplanes in the intrinsic space through (12). Note that the visible bias vector $\boldsymbol{\eta}$ is equivalent to some surface charge placed at the edge of the domain of $\mathbf{m}$ and can be incorporated into $q(\mathbf{n}, z)$. The log-likelihood of the RBM has then three terms

$$\mathcal{L}[\Theta] = -\mathbb{E}_{\hat{p}}(V[\mathbf{m}|\Theta] + \mathcal{F}^{\perp}[\mathbf{m}|\Theta]) - \log(Z[\Theta]),$$

where $\log(Z[\Theta])$ is a complex self-interaction of the charged hyperplanes among each other; $\mathbb{E}_{\hat{p}}(\mathcal{F}^{\perp}[\mathbf{m}|\Theta])$ is, in principle, small, especially if there is no transverse bias; finally,

$$\mathbb{E}_{\hat{p}}(V[\mathbf{m}|\Theta]) = \int d\mathbf{m} d\mathbf{n} dz \hat{p}(\mathbf{m})|\mathbf{n}^T\mathbf{m} - z|q(\mathbf{n}, z), \qquad (13)$$

takes the form of a repulsive Coulomb interaction between training data points represented by the empirical distribution $\hat{p}(\mathbf{m})$ and positively charged hyperplanes. It corresponds to a slight extension of the RBM model in terms of more general activation function (encompassing rectified linear unit [31], for instance, and similar to [32]), where each feature contribution in (3) comes with a non-negative weight $q_j$ to be optimized, while the features themselves defined by the pairs $(\mathbf{n}_j, \theta_j)$ are predefined. This formulation, introduced here at first in a theoretical perspective to understand the RBM, can also be used in practice when the intrinsic space is identified in advance. Then, letting $w_\beta = 0$ for $\beta > d$ results in $\mathcal{F}^{\perp}$ independent of $\Theta$ and the optimization of $\mathcal{L}[\Theta]$ [with respect to the features weights $q(\mathbf{n}, z)$] becomes convex, this "Coulomb" formulation being in the exponential family. As a result, the optimal solution can be obtained with good numerical precision thanks to a natural gradient ascent [33] following the geodesics of the Fisher metric (see Supplemental Material [29]), the complexity being $\mathcal{O}(N_f^3 + N_f^2 \times N_p^d)$ in the number of predefined features $N_f$ and of points $N_p^d$ needed to compute $Z[\Theta]$ (and its derivatives) through (5). Typically, this remains tractable for $d \leq 3$ and $N_f \leq \mathcal{O}(10^3)$ simply using a regular discretization of the feature space $(\mathbf{n}, z) \subset [-1, 1]^d$, as shown in the next section. Additionally, an even more tractable approach bypassing the computation of $Z[\Theta]$, based on a linear regression, seems plausible according to the following observations. In terms of the Coulomb charges density,

$$\rho(\mathbf{m}|\Theta) = \int d\mathbf{n} dz q(\mathbf{n}, z)\delta(\mathbf{n}^T\mathbf{m} - z), \qquad (14)$$

resulting from a distribution $q(\mathbf{n}, z)$ of uniformly charged hyperplanes, the marginal distribution of $\mathbf{m}$ reads

$$P(\mathbf{m}|\Theta) = \frac{1}{Z[\Theta]}e^{-N_v\mathcal{F}[\mathbf{m}|\Theta]},$$
$$= \frac{e^{N_v(\mathcal{S}(\mathbf{m}) + \int d\mathbf{m}'\rho(\mathbf{m}'|\Theta)K_d(|\mathbf{m}-\mathbf{m}'|) - \mathcal{F}^{\perp}[\mathbf{m}|\Theta])}}{Z[\Theta]},$$

with $K_d(|\mathbf{m} - \mathbf{m}'|)$ the inverse of the $d$-dimensional Laplacian $\nabla_d^2$ operator (see Supplemental Material [29]). Assuming for the moment that $\rho$ is not restricted to be of the specific RBM form (14), this relation can be explicitly inverted to match any smoothed version $\hat{p}_\epsilon(\mathbf{m})$ of the empirical distribution $\hat{p}(\mathbf{m})$,

$$\rho(\mathbf{m}|\Theta) = \nabla_d^2\left(\frac{1}{N_v}\log\hat{p}_\epsilon(\mathbf{m}) - \mathcal{S}[\mathbf{m}] + \mathcal{F}^{\perp}[\mathbf{m}|\Theta]\right) \qquad (15)$$

up to surface terms, provided that $\mathcal{F}^\perp$ is independent of $\rho$. Doing that leads to overfitting the data with a density of Coulomb charges concentrated on the faces of the Voronoi cells enclosing the data points (see Supplemental Material [29]). To be meaningful, this solution has to be projected on the "RBM" space, i.e., a density $\rho$ of the form (14) corresponding to a finite number of features. The fact that any distribution $\rho$ can be approximated to arbitrary precision by such a superposition of charged hyperplanes relates to the property that the RBM is a universal approximator [34]. The appropriate metric to perform such a projection is the Fisher metric [33] and this ends up being equivalent to minimizing the Kullback-Leibler divergence ($D_{\rm KL}$) between $\hat{p}(\mathbf{m})$ and $P(\mathbf{m}|\Theta)$, i.e., to maximizing the LL. Nonetheless, if we expect the optimal solution to be very close to $\hat{p}$, we may use directly the Fisher metric estimated at the empirical point $\hat{p}$, thereby turning the problem into the following linear regression:

$$\Theta^\star = \text{argmin}_{\boldsymbol{q}} \mathbb{E}_{\hat{p}}\Big[\Big|\mathcal{F}^\perp[\mathbf{m}] - \mathcal{S}[\mathbf{m}] - \sum_{j=1}^{N_h} q_j V_j[\mathbf{m}]\Big|^2\Big] \qquad (16)$$

of $\mathcal{F}^\perp[\mathbf{m}] - \mathcal{S}[\mathbf{m}]$ on the score variables $V_j[\mathbf{m}] \overset{\text{def}}{=} \{\partial \log[P(\mathbf{m}|\Theta)]/\partial q_j\}$ conjugate to $q_j$ (see Supplemental Material [29]).

*Study cases.*—To illustrate these statements, first consider a dataset supported by a 1D subspace given by the vector $u_i = 1/\sqrt{N_v}$ with unbiased fluctuations along other directions. A rank one $W = w_1 u^1 v^{1T}$ is assumed since we expect transverse modes to vanish from the linear stability analysis of the training given in [28]. The relation (4) reduces then to the magnetization $m = \tanh(\mu)$ along $u$ leading in the Coulomb formulation to

$$\mathcal{F}[m|\boldsymbol{q}] = h(m) - \sum_{j=0}^{N_h} q_j |m - z_j|, \qquad (q_j \geq 0),$$

with $h(m) = \frac{1}{2}(1 \pm m)\log(1 \pm m)$. The natural gradient ascent of the LL yields an optimal solution as the one shown on Fig. 2. As is manifest on Fig. 2, the result is the linear regression (16) of $\mathcal{F}^\perp[\mathbf{m}] - \mathcal{S}[\mathbf{m}] = h(m)$ in terms of a piecewise linear function, where the break points correspond to the locations $z_j$ of the relevant features and $q_j$ the corresponding break of slope at these points. This involves, however, an implicit regularization, which will be studied elsewhere, in order to maintain the regions free of data below $h(m)$ in order to stay away from first-order transitions where the local Fisher metric would cease to be a meaningful approximation to the $D_{\rm KL}$. As a 2D example, we consider data concentrated in the subspace spanned by the vectors $u_i^1 = 1/\sqrt{N_v}$ and $u_i^2 = (-1)^i/\sqrt{N_v}$ with irrelevant transverse fluctuations, hence assuming now $W = w_1 u^1 v^{1T} + w_2 u^2 v^{2T}$. We have then a finite magnetization $(m_1, m_2)$ along each direction and the free energy considered in the Coulomb formulation reads

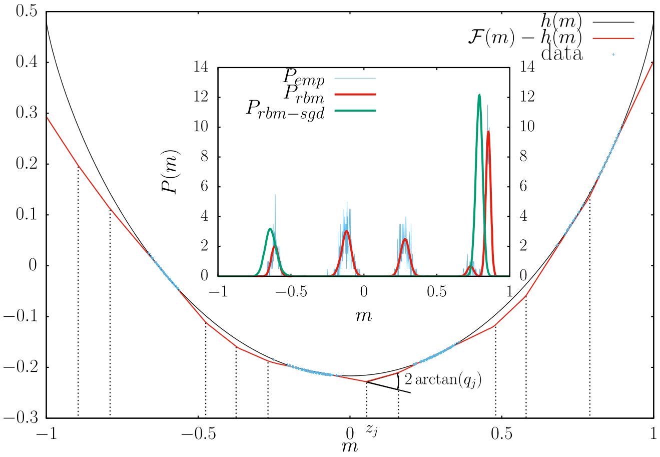

FIG. 2. 1D intrinsic data ($N_v = 10^3$) with five clusters solved with $N_h = 20$ predefined features thanks to a natural gradient ascent of the LL. Dotted lines indicate location of features with nonvanishing weights $q_j$. The feature contributions $\mathcal{F}(m) - h(m)$ to the free energy are seen to regress $h(m)$ on the data. The resulting distribution is shown (red) on the inset with the empirical training distribution (blue) and the *failed* result of a standard RBM training (green).

$$\mathcal{F}[\mathbf{m}|\Theta] = \frac{1}{2}[h(m^+) + h(m^-)] \\ - \sum_{j=1}^{N_h} q_j |m_1 \cos(\omega_j) + m_2 \sin(\omega_j) - z_j|,$$

where $m^\pm = m_1 \pm m_2 \in [-1, 1]$ and $\omega_j \in [0, \pi[$ are the angles made by the charged lines with the $m_2$ axis.

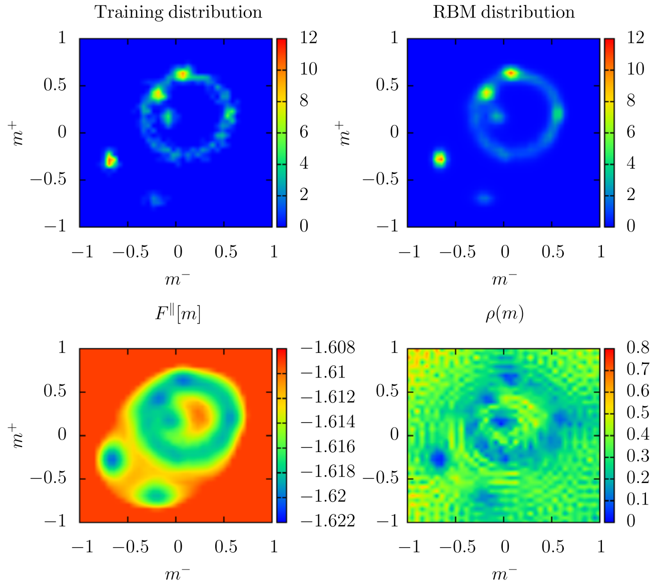

FIG. 3. 2D intrinsic dataset ($N_v = 10^3$) with six pointlike clusters and a circular one (upper left) and corresponding RBM density (upper right) found with $N_h = 900$ predefined features, along with its free energy landscape (bottom left) and Coulomb charges distribution (bottom right).

The result of the natural gradient ascent of the LL is shown on Fig. 3. Here a large number of features $(\omega_j, z_j) \in [0, \pi] \times [-1, 1]$ have been predefined on a regular lattice in order to obtain a continuous charge distribution and a smooth free energy landscape (see more details in the Supplemental Material [29]). Finally, in both study cases, the standard RBM training fails for two distinct reasons unveiled by the Coulomb picture (see Supplemental Material [29]): (i) the Gibbs sampling is plagued by the presence of first-order phase transitions with respect to an annealing temperature, and (ii) the charged hyperplanes get easily trapped by Coulomb barriers formed by the clusters of data, a pitfall bypassed by the convex Coulomb relaxation.

*Discussion.*—The physical picture of the RBM emerging here, in addition to identifying and disentangling via Eqs. (11) and (12) the role played by some key factors, underlines the importance of two distinct aspects of learning a high dimensional distribution: the ordered part corresponding to global statistical patterns and the fluctuations around these patterns encoding possibly short range correlations or corresponding to noise. Under a flat intrinsic space hypothesis, our formalism decouples them and gives indications of how to learn them separately in order to obtain high quality models that are needed in scientific applications, when default RBM algorithms are thwarted by low dimensional global patterns as we see in our experiments. Among many possible developments, we foresee that the Coulomb convex relaxation could be used to fine-tune some otherwise poorly trained RBM and opens the intriguing possibility of tackling unsupervised learning via regularized linear regressions.

A. D. was supported by the Comunidad de Madrid and the Complutense University of Madrid (Spain) through the Atracción de Talento program (Ref. 2019-T1/TIC-13298).

[*]Corresponding author.
cyril.furtlehner@inria.fr

[1] G. Torlai and R. G. Melko, Learning thermodynamics with Boltzmann machines, Phys. Rev. B **94**, 165134(2016).

[2] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[3] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Restricted Boltzmann machine learning for solving strongly correlated quantum systems, Phys. Rev. B **96**, 205152 (2017).

[4] R. G. Melko, G. Carleo, and J. Carrasquilla, Restricted Boltzmann machines in quantum physics, Nat. Phys. **15**, 887 (2019).

[5] J. Tubiana, S. Cocco, and R. Monasson, Learning protein constitutive motifs from sequence data, eLife **8**, e39397 (2019).

[6] P. Smolensky, in *Parallel Distributed Processing*, edited by D. Rumelhart and J. McLelland (MIT Press, Cambridge, MA, 1986), Vol. 1, Chap. 6, pp. 194–281.

[7] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Neural Information Processing Systems* (2014), p. 2672.

[9] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, in *International Conference on Learning Representations* (2014).

[10] R. Salakhutdinov and G. Hinton, Deep Boltzmann machines, in *Artificial Intelligence and Statistics* (PMLR, Florida, 2009), p. 448.

[11] R. D. Hjelm, V. D. Calhoun, R. Salakhutdinov, E. A. Allen, T. Adali, and S. M. Plis, Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks, NeuroImage **96**, 245 (2014).

[12] X. Hu, H. Huang, B. Peng, J. Han, N. Liu, J. Lv, L. Guo, C. Guo, and T. Liu, Latent source mining in fmri via restricted Boltzmann machine, Hum. Brain Mapp. **39**, 2368 (2018).

[13] B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, C. Furtlehner, L. Pagani, and F. Jay, Creating artificial human genomes using generative neural networks, PLoS Genet. **17**, e1009303 (2021).

[14] G. E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Comput. **14**, 1771 (2002).

[15] T. Tieleman, Training restricted Boltzmann machines using approximations to the likelihood gradient, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (Association for Computing Machinery (ACM), Helsinki, 2008), p. 1064–1071.

[16] A. Fischer and C. Igel, Training restricted Boltzmann machines: An introduction, Pattern Recognit. **47**, 25 (2014).

[17] M. Gabrié, E. W. Tramel, and F. Krzakala, Training restricted Boltzmann machine via the TAP free energy, Adv. Neural Inf. Process. Syst. **28**, 640 (2015).

[18] H. Huang and T. Toyoizumi, Advanced mean-field theory of the restricted Boltzmann machine, Phys. Rev. E **91**, 050101(R) (2015).

[19] C. Takahashi and M. Yasuda, Mean-field inference in Gaussian restricted Boltzmann machine, J. Phys. Soc. Jpn. **85**, 034001 (2016).

[20] M. Mézard, Mean-field message-passing equations in the Hopfield model and its generalizations, Phys. Rev. E **95**, 022117 (2017).

[21] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, On the equivalence of Hopfield networks and Boltzmann machines, Neural Netw. **34**, 1 (2012).

[22] A. Barra, G. Genovese, P. Sollich, and D. Tantari, Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors, Phys. Rev. E **97**, 022310 (2018).

[23] A. Barra, G. Genovese, P. Sollich, and D. Tantari, Phase transitions in restricted Boltzmann machines with generic priors, Phys. Rev. E **96**, 042156 (2017).

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

[25] R. Monasson and J. Tubiana, Emergence of Compositional Representations in Restricted Boltzmann Machines, Phys. Rev. Lett. **118**, 138301 (2017).

[26] A. Decelle and C. Furtlehner, Restricted Boltzmann machine, recent advances and mean-field theory, Chin. Phys. B **30**, 040202 (2020).

[27] A. Decelle, G. Fissore, and C. Furtlehner, Spectral dynamics of learning in restricted Boltzmann machines, Europhys. Lett. **119**, 60001 (2017).

[28] A. Decelle, G. Fissore, and C. Furtlehner, Thermodynamics of restricted Boltzmann machines and related learning dynamics, J. Stat. Phys. **172**, 1576 (2018).

[29] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.158303 for technical details or additional numerical results.

[30] Note that, practically speaking, we use finite $N_v$ estimates of $\Phi$ and $m_\alpha$ so that the preceding relation is in fact valid up to some $\mathcal{O}(1/\sqrt{N_v})$ corrections with respect to a limit defined by some hypothetical $p_{\boldsymbol{u}}$ when $N_v \to \infty$.

[31] V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *ICML '10* (Omnipress, Haifa, 2010), pp. 807–814.

[32] W. Ping, Q. Liu, and A. T. Ihler, Learning infinite RBMs with Frank-Wolfe, in *Neural Information Processing Systems* (2016), Vol. 29.

[33] S.-I. Amari, Natural gradient works efficiently in learning, Neural Comput. **10**, 251 (1998).

[34] N. Le Roux and Y. Bengio, Representational power of restricted Boltzmann machines and deep belief networks, Neural Comput. **20**, 1631 (2008).