HAL
open science

Exact Training of Restricted Boltzmann Machines on Intrinsically Low Dimensional Data

A Decelle, Cyril Furtlehner

To cite this version:

A Decelle, Cyril Furtlehner. Exact Training of Restricted Boltzmann Machines on Intrinsically Low Dimensional Data. Physical Review Letters, 2021, 10.1103/physrevlett.127.158303 . hal-03432350

HAL Id: hal-03432350

https://inria.hal.science/hal-03432350v1

Submitted on 17 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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.

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):

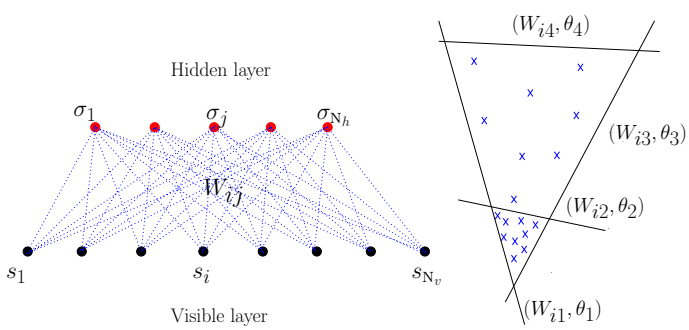

FIG. 1. Bipartite structure of the RBM (left). Hyperplanes defined by the weight vectors and bias associated with each hidden variable can delimit fixed density regions in input space (right).

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

The former noted $\mathbf{s} = \{s_i, i = 1 \dots N_v\}$ correspond to explicit representations of the data while the latter noted $\boldsymbol{\sigma} = \{\sigma_j, j = 1 \dots N_h\}$ 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 Θ 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

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

with $\langle\rangle_{\text{Data}}$ denoting the average over training data. This is a nontrivial optimization problem in two respects: it is nonconvex and the loss function $-\mathcal{L}[\Theta]$ is difficult to estimate because $\log\big(Z[\Theta]\big)$ 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 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

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

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

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

For a RBM trained on some data, d would represent their intrinsic dimension at least locally. The corresponding modes u_i^α 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 $\mathbf{s} \longrightarrow (\mathbf{m}, \mathbf{s}^\perp)$, 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 $\mathcal{S}[\mathbf{m}] = \frac{\mathcal{N}[\mathbf{m}]}{N_v}\log(2)$ under these constraints. Thanks to a large deviation argument $\mathcal{S}[\mathbf{m}]$ is the Legendre transform of (see SM, Appendix A)

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

with $\boldsymbol{\mu}[\mathbf{m}]$ given implicitly by the constraints [29]

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

Given a condensed magnetization vector $\mathbf{m}$, there remains $\mathcal{N}[\mathbf{m}]$ interacting degrees of freedom represented by spin variables denoted $\{s_1^\perp, \dots, s_{\mathcal{N}[\mathbf{m}]}^\perp\}$. 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 SM, Appendix B):

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

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

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

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

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 paramagnetic-like state of order defined for each $\mathbf{m}$ (see SM, Appendix C). 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 $\mathcal{H}_{\text{eff}} = 0$. 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 re-written as (See SM Appendix D)

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

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

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

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

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

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

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 ν_j^{-1}. At this point let us remark that the w_α control through (11) both the strength of the Coulomb interaction via (9,10) and the charged hyperplanes thickness; the right singular vectors projections v_j^α 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

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

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

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

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 RELU [30] for instance and similar to [31]), 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 Θ and the optimization of $\mathcal{L}[\Theta]$ (w.r.t. 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 [32] following the geodesics of the Fisher metric (See SM, Appendix G), 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

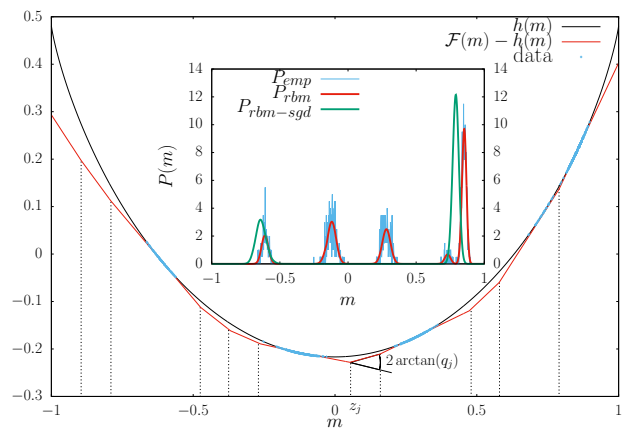

FIG. 2. 1-d intrinsic data ($N_v = 10^3$) with 5 clusters solved with $N_h = 20$ predefined features thanks to a natural gradient ascent of the LL. Dotted lines indicate location of features with nonvanishing weights q_j. The feature contributions $\mathcal{F}(m) - h(m)$ to the free energy are seen to 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).

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

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

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

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

with $K_d(|\mathbf{m}-\mathbf{m}'|)$ the inverse of the d-dimensional Laplacian ∇_d^2 operator (See SM, Appendix D). Assuming for the moment that ρ 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 $\hat{p}(\mathbf{m})$:

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

up to surface terms, provided that $\mathcal{F}^{\perp}$ is independent of ρ. Doing that leads to overfit the data with a density of Coulomb charges concentrated on the faces of the Voronoi cells enclosing the data points (see SM, Appendix E). To be meaningful this solution has to be projected on the "RBM" space, i.e. a density ρ of the form (14) corresponding to a finite number of features. The fact that any distribution ρ can be approximated to arbitrary precision by such a superposition of charged hyperplanes relates to the property that the RBM is a universal approximator [33]. The appropriate metric to perform such a projection is the Fisher metric [32] and this ends up being equivalent to minimizing the Kullback-Leibler divergence (D_{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

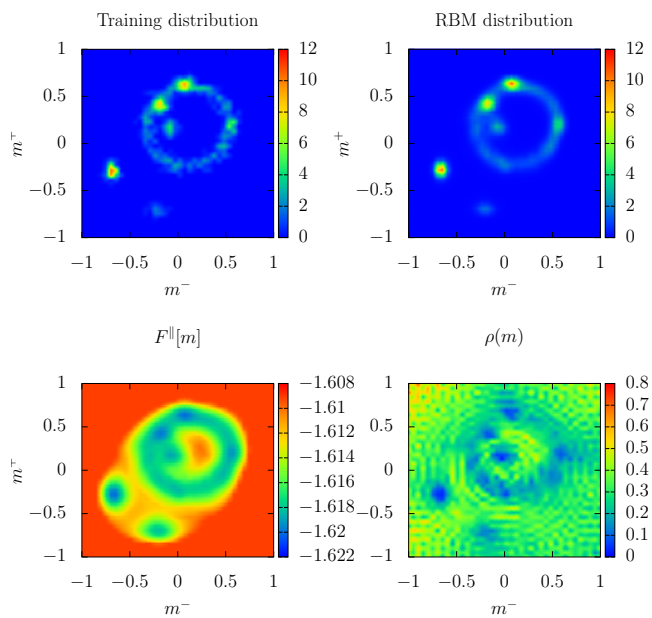

FIG. 3. 2-d intrinsic dataset ($N_v = 10^3$) with 6 point-like clusters and a circular one (upper left) and corresponding RBM density (upper right) found with $N_h = 900$ predefined features, along with its free energy landscape (bottom left) and Coulomb charges distribution (bottom right)

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

of $\mathcal{F}^{\perp}[\mathbf{m}] - \mathcal{S}[\mathbf{m}]$ on the score variables $V_j[\mathbf{m}] \stackrel{\text{def}}{=} \frac{\partial \log(P(\mathbf{m}|\Theta))}{\partial q_j}$ conjugate to q_j (see SM, Appendix F).

Study cases. — To illustrate these statements first consider a dataset supported by a 1-d 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 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_{KL}. As a 2-d 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

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

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. 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 SM, Appendix G). Finally, in both study cases the standard RBM training fails for two distinct reasons unveiled by the Coulomb picture (see SM, Appendix G): (i) the Gibbs sampling is plagued by the presence of first-order phase transitions with respect to an annealing temperature; (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. (11,12) 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.

Acknowledgments 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).

[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(6325):602–606, 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.*, 15:887–892, 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: Volume 1 by D. Rumelhart and J. McLelland*, chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. 194-281. MIT Press, 1986.

[7] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 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 *NIPS*, pages 2672–2680, 2014.

[9] D.P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[10] R. Salakhutdinov and G. Hinton. Deep Boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

[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*, 96:245–260, 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. *Human brain mapping*, 39(6):2368–2380, 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 Genetics*, 17:1–22, 02 2021.

[14] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14:1771–1800, 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, pages 1064–1071, New York, NY, USA, 2008. ACM.

[16] A. Fischer and C. Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

[17] M. Gabrié, E.W. Tramel, and F. Krzakala. Training restricted Boltzmann machine via the TAP free energy. In *Advances in Neural Information Processing Systems 28*, pages 640–648. 2015.

[18] H. Huang and T. Toyoizumi. Advanced mean-field theory of the restricted Boltzmann machine. *Phys. Rev. E*, 91(5):050101, 2015.

[19] C. Takahashi and M. Yasuda. Mean-field inference in Gaussian restricted Boltzmann machine. *Journal of the Physical Society of Japan*, 85(3):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 Networks*, 34:1–9, 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(4):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. Let.*, 118:138301, 2017.

[26] A. Decelle and C. Furtlehner. Restricted Boltzmann machine, recent advances and mean-field theory. *Chinese Physics B*, 2020.

[27] A. Decelle, G. Fissore, and C. Furtlehner. Spectral dynamics of learning in restricted Boltzmann machines. *EPL*, 119(6):60001, 2017.

[28] A. Decelle, G. Fissore, and C. Furtlehner. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *J.Stat.Phys.*, 172(18):1576–1608, 2018.

[29] Note that practically speaking we use finite N_v estimates of Φ and m_α so that the preceding relation is in fact valid up to some $\mathcal{O}(1/\sqrt{N_v})$ corrections w.r.t. limit defined by some hypothetical $p_{\boldsymbol{u}}$ when $N_v \to \infty$.

[30] V. Nair and G.E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *ICML '10*, pages 807–814, 2010.

[31] W. Ping, Q. Liu, and A.T. Ihler. Learning infinite RBMs with Frank-Wolfe. In *NIPS*, volume 29, 2016.

[32] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
[33] N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.
[34] H. Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478:1–69, 2009.

Appendix A: Canonical ensemble with magnetization constraints

The decomposition of the vector of visible variables $\mathbf{s}$ on the left singular basis

$$s_\alpha \stackrel{\text{def}}{=} \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} s_i u_i^\alpha, \tag{A1}$$

coincides with m_α for $\alpha = 1, \dots d$ by definition of the magnetization constraints. We look for a change of variables $\mathbf{s} \longrightarrow (\mathbf{m}, \mathbf{s}^\perp)$ where the original spin variables are replaced by the set of $\mathbf{m} = \{m_\alpha, \alpha = 1, \dots d\}$ and $\mathcal{N}[\mathbf{m}]$ transverse spin variables. Let us denote by $\mathbb{E}_{\boldsymbol{s}\sim\mathbb{U}}$ the expectation taken when the original spin variables are iid, $s_i \sim \mathbb{U}_{\{-1,1\}}$. The change of measure is made by looking at the prior distribution over the original spin variables:

$$P_{\text{prior}}[\mathbf{s}] = \frac{1}{2^{N_v}} = P_{\text{prior}}[\mathbf{s}^\perp|\mathbf{m}] P_{\text{prior}}[\mathbf{m}],$$

where

$$P_{\text{prior}}[\mathbf{m}] = \mathbb{E}_{\boldsymbol{s}\sim\mathbb{U}}\Big[\prod_{\alpha=1}^{d} \delta(s_\alpha - m_\alpha)\Big] = e^{N_v(\mathcal{S}[\mathbf{m}] - \log 2)} \tag{A2}$$

represents the density of states (normalized to one) associated with the magnetization constraints $\mathbf{m}$, $\mathcal{S}[\mathbf{m}]$ the configuration entropy associated with these magnetizations and

$$P_{\text{prior}}[\mathbf{s}^\perp|\mathbf{m}] \stackrel{\text{def}}{=} \frac{1}{2^{\mathcal{N}[\mathbf{m}]}},$$

with $\mathcal{N}[\mathbf{m}] = N_v S[\mathbf{m}]/\log(2)$ representing the remaining number of degrees of freedom $\mathbf{s}^\perp$ taken out of the N_v initial ones. Note that there is a formal difficulty here because the size of the transverse variables vector $\mathbf{s}^\perp$ depends explicitly on $\mathbf{m}$. This however is not really a problem if consider $\mathbf{s}^\perp$ to be a vector of size N_v where the last $N_v - \mathcal{N}(\mathbf{m})$ bits are frozen arbitrarily to 1, which is done in practice by defining the prior distribution

$$P_{\text{prior}}[\mathbf{s}^\perp|\mathbf{m}] \stackrel{\text{def}}{=} \frac{1}{2^{\mathcal{N}[\mathbf{m}]}} \prod_{\ell=\mathcal{N}[\mathbf{m}]+1}^{N_v} \delta(s_\ell^\perp - 1).$$

To avoid additional burden on the notations we keep this as implicit and $\mathbf{s}^\perp$ always refers to the set of non-frozen variables. We want here to determine $\mathcal{S}[\mathbf{m}]$ from (A2). We have

$$\mathbb{E}_{\boldsymbol{s}\sim\mathbb{U}}[s_\alpha] = 0 \qquad \text{and} \qquad \mathbb{E}_{\boldsymbol{s}\sim\mathbb{U}}[s_\alpha s_\beta] = \frac{1}{N_v}\delta_{\alpha\beta},$$

the second relation resulting from the orthogonality of the $\mathbf{u}^\alpha$ vectors. As a result, for large N_v we have

$$P_{\text{prior}}[\mathbf{m}] = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{N_v}{2}\sum_{\alpha=1}^{d} m_\alpha^2\Big). \tag{A3}$$

This is valid as long as the magnetization are not too large ($m_\alpha = \mathcal{O}(1/\sqrt{N_v})$). To study the regime where modes condense, i.e. when $m_\alpha = \mathcal{O}(1)$, we have to resort to large deviations estimations [34]. With d assumed to be $\mathcal{O}(1)$, as $N_v \to \infty$ we expect in this regime a behaviour of the form

$$P_{\text{prior}}[\mathbf{m}] \asymp e^{-N_v \mathcal{I}[\mathbf{m}]},$$

where $\mathcal{I}[\mathbf{m}]$ called the rate function, has 0 as minimum value and can be determined in the present situation thanks to the Gärtner-Ellis theorem from the moment generating function of $P_{\text{prior}}[\mathbf{m}]$. Denoting by $\boldsymbol{\mu} = \mathcal{O}(1)$ a conjugate d-dimensional vector and assuming that we can make sense of the following limit

$$\Phi[\boldsymbol{\mu}] \stackrel{\text{def}}{=} \lim_{N_v\to\infty} \frac{1}{N_v} \log\Big(\mathbb{E}_{\boldsymbol{s}\sim\mathbb{U}}\Big[e^{N_v \sum_{\alpha=1}^{d} m_\alpha(\boldsymbol{s})\mu_\alpha}\Big]\Big),$$

$$= \lim_{N_v\to\infty} \frac{1}{N_v} \sum_i \log\cosh\Big(\sqrt{N_v}\sum_{\alpha=1}^{d} u_i^\alpha \mu_\alpha\Big),$$

$\mathcal{I}[\mathbf{m}]$ is then simply given by the Legendre-Fenchel transform of Φ:

$$\mathcal{I}[\mathbf{m}] = \boldsymbol{m}\boldsymbol{\mu}[\mathbf{m}]^T - \Phi\big[\boldsymbol{\mu}[\mathbf{m}]\big],$$

with $\boldsymbol{\mu}[\mathbf{m}]$ implicitly given by (in principle when $N_v \to \infty$)

$$m_\alpha = \lim_{N_v\to\infty} \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} u_i^\alpha \tanh\Big(\sqrt{N_v}\sum_{\beta=1}^{d} u_i^\beta \mu_\beta\Big),$$

$$= \mathbb{E}_{\boldsymbol{u}\sim p_{\boldsymbol{u}}}\Big[u^\alpha \tanh\Big(\sum_{\beta=1}^{d} u^\beta \mu_\beta\Big)\Big] \tag{A4}$$

where we assume in the last equality some limit $p_{\boldsymbol{u}}$ of the joint empirical distribution of $u^\alpha = \sqrt{N_v} u_i^\alpha$ when $N_v \to \infty$. From the small $\mathbf{m}$ behaviour given in (A3) we finally have determined the configuration entropy as

$$\mathcal{S}[\mathbf{m}] = -\mathcal{I}[\mathbf{m}] + \log(2) + \frac{d}{2N_v}\log(2\pi),$$

$$= \Phi[\boldsymbol{\mu}[\mathbf{m}]] - \mathbf{m}^T\boldsymbol{\mu}[\mathbf{m}] + \log(2) + \mathcal{O}\Big(\frac{1}{N_v}\Big).$$

Note that practically speaking we use finite N_v estimates of Φ and m_α so that the preceding relation is in fact valid up to some $\mathcal{O}(1/\sqrt{N_v})$ corrections w.r.t. limit defined by some hypothetical $p_{\boldsymbol{u}}$ when $N_v \to \infty$.

Appendix B: Longitudinal and transverse free energy

In order to disentangle the contributions of the collective modes materialized by some magnetization $\mathbf{m}$ along some directions $\mathbf{u}^\alpha, \alpha = 1, \dots d$ from the noise corresponding to the fluctuations of the transverse variables $\mathbf{s}^\perp$ we assume first to be able to rewrite the Hamiltonian corresponding to the visible distribution (3)

$$P[\mathbf{s}|\Theta] = \frac{e^{-\mathcal{H}[\mathbf{s}|\Theta]}}{Z[\Theta]},$$

in terms of the new degrees of freedoms

$$\mathcal{H}[\mathbf{s}|\Theta] = \mathcal{H}[\mathbf{m}, \mathbf{s}^\perp|\Theta],$$

such that the joint distribution takes the form

$$P[\mathbf{m}, \mathbf{s}^\perp|\Theta] = \frac{e^{-\mathcal{H}[\mathbf{m},\mathbf{s}^\perp|\Theta]}}{Z[\Theta]}.$$

This in turn can be written

$$P[\mathbf{m}, \mathbf{s}^\perp|\Theta] = P[\mathbf{s}^\perp|\mathbf{m}, \Theta]P[\mathbf{m}|\Theta], \tag{B1}$$

with

$$P[\mathbf{m}|\Theta] = \sum_{\mathbf{s}} P[\mathbf{s}] \prod_{\alpha=1}^{d} \delta(s_\alpha - m_\alpha) \stackrel{\text{def}}{=} \frac{e^{-N_v \mathcal{F}[\mathbf{m}|\Theta]}}{Z[\Theta]},$$

after introducing the canonical free energy $\mathcal{F}[\mathbf{m}|\Theta]$. Let us denote by

$$\mathcal{H}[\mathbf{s}^\perp|\mathbf{m}, \Theta] \stackrel{\text{def}}{=} \mathcal{H}[\mathbf{m}, \mathbf{s}^\perp|\Theta] - \mathcal{H}_0[\mathbf{m}|\Theta], \tag{B2}$$

the "conditional" Hamiltonian, where $\mathcal{H}_0[\mathbf{m}|\Theta]$ is at this point an arbitrary function of $\mathbf{m}$ independent of $\mathbf{s}^\perp$. We chose it as to contain only contributions from the longitudinal magnetization, i.e. coincides with the constant part of $\mathcal{H}[\mathbf{m}, \mathbf{s}^\perp|\Theta]$ w.r.t. $\mathbf{s}^\perp$ when neglecting the singular values w_β of W and the $\mathcal{O}\Big(1/\sqrt{N_v}\Big)$ residual transverse magnetizations m_β for $\beta > d$ resulting from the constraints (see next Section). Rewriting the Hamiltonian in terms of the SVD components s_α and η_α respectively of the visible variables and biases

$$\begin{aligned}\mathcal{H}[\mathbf{s}|\Theta] &= \sum_{i=1}^{N_v} \eta_i s_i - \sum_{j=1}^{N_h} \log\cosh\Big(\sum_{i=1}^{N_v} W_{ij} s_i - \theta_j\Big) \\ &= N_v \sum_{\alpha=1}^{N_v} \eta_\alpha s_\alpha - \sum_{j=1}^{N_h} \log\cosh\Big(\sqrt{N_v} \sum_{\alpha=1}^{N_v} w_\alpha s_\alpha v_j^\alpha - \theta_j\Big),\end{aligned}$$

this leads to the definition

$$\mathcal{H}_0[\mathbf{m}|\Theta] = N_v\big(\sum_{\alpha=1}^{d} \eta_\alpha m_\alpha - V(\mathbf{m}|\Theta)\big), \tag{B3}$$

with

$$V(\mathbf{m}|\Theta)) \stackrel{\text{def}}{=} \frac{1}{N_v} \sum_{j=1}^{N_h} \log\cosh\Big(\sqrt{N_v} \sum_{\alpha=1}^{d} w_\alpha m_\alpha v_j^\alpha - \theta_j\Big).$$

In terms of $\mathcal{H}_0[\mathbf{m}|\Theta]$ the free energy reads

$$\begin{aligned}\mathcal{F}[\mathbf{m}|\Theta] &= \frac{1}{N_v}\Big[\mathcal{H}_0[\mathbf{m}|\Theta] - \log\Big(\sum_{\mathbf{s}^\perp} e^{-\mathcal{H}[\mathbf{s}^\perp|\mathbf{m},\Theta]}\Big)\Big], \\ &= \frac{1}{N_v}\Big[\mathcal{H}_0[\mathbf{m}|\Theta] - \mathcal{S}[\mathbf{m}] - \log\Big(\frac{\sum_{\mathbf{s}^\perp} e^{-\mathcal{H}[\mathbf{s}^\perp|\mathbf{m},\Theta]}}{2^{\mathcal{N}[\mathbf{m}]}}\Big)\Big] \\ &= \mathcal{F}^{\parallel}[\mathbf{m}|\Theta] + \mathcal{F}^{\perp}[\mathbf{m}|\Theta]\end{aligned}$$

where we have introduced respectively the longitudinal and transverse free energy:

$$\mathcal{F}^{\parallel}[\mathbf{m}|\Theta] \stackrel{\text{def}}{=} \frac{1}{N_v}(\mathcal{H}_0[\mathbf{m}|\Theta] - \mathcal{S}[\mathbf{m}]),$$

$$\mathcal{F}^{\perp}[\mathbf{m}|\Theta] \stackrel{\text{def}}{=} -\frac{1}{N_v} \log\Big(\frac{1}{2^{\mathcal{N}[\mathbf{m}]}} \sum_{\mathbf{s}^\perp} e^{-\mathcal{H}[\mathbf{s}^\perp|\mathbf{m},\Theta]}\Big).$$

The longitudinal free energy of the system is the free energy of the system for a given magnetization $\mathbf{m}$ when the transverse magnetization and interactions among the $s_\ell^\perp$ are neglected. These indeed are expected to be small by definition of the intrinsic space. Vanishing interactions corresponds to $\mathcal{H}\big[\mathbf{s}^\perp|\mathbf{m}, \Theta\big] = 0$, in which case $\mathcal{F}^{\parallel}[\mathbf{m}|\Theta]$ coincides with $\mathcal{F}[\mathbf{m}|\Theta]$. Non-vanishing interactions are accounted for by the transverse free energy.

Appendix C: Effective Hamiltonian

To enter further into the description of transverse fluctuations we need to specify the transverse degrees of freedom $\mathbf{s}^\perp$ and the way they interact through $\mathcal{H}[\mathbf{s}^\perp|\mathbf{m}, \Theta]$ in the form of an effective Hamiltonian. For $\beta > d$ the components $\mathbf{s}_\beta$ given in (A1) are stochastic variables and we denote them by $s_\beta[\mathbf{s}^\perp|\mathbf{m}]$, i.e. a mapping to be defined of transverse variables to transverse projections given some magnetization $\mathbf{m}$. This mapping cannot be determined exactly but since we look for an effective theory we consider a linear map i.e. of the form

$$s_\beta[\mathbf{s}^\perp|\mathbf{m}] = m_\beta[\boldsymbol{\mu}] + \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} c_\beta^\ell s_\ell^\perp, \qquad \beta = d+1, \dots N_v, \tag{C1}$$

where the prior of these new variables is to be iid with $s_\ell^\perp \sim \mathbb{U}_{\{-1,1\}}$. Under the magnetization constraints *only* we have (see Section A)

$$\mathbb{E}[s_i|\mathbf{m}] = m_i[\boldsymbol{\mu}]$$

$$\mathrm{Cov}(s_i, s_j|\mathbf{m}) = (1 - m_i^2[\boldsymbol{\mu}])\delta_{ij} + \mathcal{O}\Big(\frac{1}{N_v}\Big)$$

with

$$m_i[\boldsymbol{\mu}] = \tanh\Big(\sqrt{N_v}\sum_{\alpha=1}^{d}\mu_\alpha u_i^\alpha\Big).$$

As a result $m_\beta[\boldsymbol{\mu}]$ (for $\beta > d$) represents the prior bias resulting from the magnetizations constraints (A4)

$$m_\beta[\boldsymbol{\mu}] = \frac{1}{\sqrt{N_v}}\sum_{i=1}^{N_v} u_i^\beta \tanh\Big(\sqrt{N_v}\sum_{\alpha=1}^{d}\mu_\alpha u_i^\alpha\Big),\ \forall\beta > d \tag{C2}$$

as a function of $\{\mu_\alpha, \alpha = 1,\dots d\}$ solution to equation (A4) and are $\mathcal{O}(1/\sqrt{N_v})$. The second set of constraints comes from the set of prior covariances between the $s_\beta[\mathbf{s}^\perp|\mathbf{m}]$ for $\beta > d$ that have to be maintained to properly account for the transverse degrees of freedom. These are diagonal at leading order:

$$\mathrm{Cov}(s_\beta[\mathbf{s}^\perp|\mathbf{m}], s_\gamma[\mathbf{s}^\perp|\mathbf{m}]) = \frac{1}{N_v}\sum_{i=1}^{N_v} {u_i^\beta}^2(1 - m_i^2[\boldsymbol{\mu}])\delta_{\beta\gamma} + \mathcal{O}\Big(\frac{1}{N_v^2}\Big). \tag{C3}$$

A simple way to maintain at best these covariances in the new representation is to associate a binary variable $s_\ell^\perp$ with the first $\mathcal{N}[\mathbf{m}]$ principal axes of the previous covariance matrix neglecting the remainder. Since the covariance resulting from (C1) reads

$$\mathrm{Cov}(s_\beta[\mathbf{s}^\perp|\mathbf{m}], s_\gamma[\mathbf{s}^\perp|\mathbf{m}]) = \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} c_\beta^\ell c_\gamma^\ell,$$

the vector $\boldsymbol{c}_\ell$ can be chosen as a principal axes of the covariance matrix (C3) normalized to the standard deviation along the same axes which are $\mathcal{O}\Big(1/\sqrt{N_v}\Big)$ so that c_ℓ^β coefficients are $\mathcal{O}\Big(1/N_v\Big)$. Now that the transverse variables are unambiguously defined we can obtained their effective Hamiltonian by expanding $\mathcal{H}[\mathbf{s}^\perp|\mathbf{m}]$ defined in (B2) at second order in $s^\perp$. From the definition (B2,B3) there is a zero order term $\mathcal{F}_0^\perp[\mathbf{m}|\Theta] = \mathcal{O}(1/\sqrt{N_v})$ with convoluted expression which we give only the first terms, contributing to the transverse free energy and the first and second order terms corresponds to a conventional Hamiltonian of a disordered Ising model:

$$\begin{aligned}\mathcal{H}[\mathbf{s}^\perp|\mathbf{m},\Theta] &\approx \mathcal{H}_{\mathrm{eff}}\big[\mathbf{s}^\perp|\mathbf{m},\Theta\big] \\ &= N_v\mathcal{F}_0^\perp[\mathbf{m}|\Theta] \\ &+ \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]}\eta_\ell^\perp[\mathbf{m},\Theta]s_\ell^\perp + \sum_{\ell,\ell'=1}^{\mathcal{N}[\mathbf{m}]} W_{\ell\ell'}^\perp[\mathbf{m},\Theta]s_\ell^\perp s_{\ell'}^\perp.\end{aligned}$$

with

$$\mathcal{F}_0^\perp[\mathbf{m}|\Theta] = \sum_{\beta=d+1}^{N_v}(\eta_\beta + w_\beta\bar{m}_\beta)m_\beta[\mu]$$

$$\eta_\ell^\perp[\mathbf{m},\Theta] = N_v\sum_{\beta=d+1}^{N_v} c_\beta^\ell(\eta_\beta - w_\beta\bar{m}_\beta)$$

$$W_{\ell,\ell'}^\perp[\mathbf{m},\Theta] = -N_v\sum_{j=1}^{N_h}(1-\bar{m}_j^2)\sum_{\beta,\gamma=d+1}^{N_v} w_\beta w_\gamma c_\beta^\ell c_\gamma^{\ell'} v_j^\beta v_j^\gamma,$$

after introducing the notations:

$$\bar{m}_j = \tanh\Big(\sqrt{N_v}\big(\sum_{\alpha=1}^{d} w_\alpha m_\alpha v_j^\alpha + \sum_{\beta=d+1}^{N_v} w_\beta m_\beta[\boldsymbol{\mu}]v_j^\beta\big) - \theta_j\Big),$$

$$\bar{m}_\beta = \frac{1}{\sqrt{N_v}}\sum_{j=1}^{N_h}\bar{m}_j v_j^\beta.$$

$\eta_\ell^\perp[\mathbf{m},\Theta]$ is potentially $\mathcal{O}(1)$ while $W_{\ell,\ell'}^\perp[\mathbf{m},\Theta]$ is $\mathcal{O}\Big(\frac{1}{\sqrt{N_v}}\Big)$.

To make connection with data, i.e. given a configuration $\mathbf{s}$ with magnetization $m_\alpha, \alpha = 1,\dots d$, from which are extracted the transverse magnetization $\{m_\beta[\boldsymbol{\mu}], \beta = d+1,\dots N_v$ by solving the equations (A4), the $\mathbf{s}^\perp$ are constructed as follows. First let for each $\ell = 1,\dots,\mathcal{N}[\mathbf{m}]$

$$m_\ell^\perp[\mathbf{s}] = \sum_{\beta=d+1}^{N_v}(s_\beta - m_\beta[\boldsymbol{\mu}])u_\beta^{\perp\ell}$$

(thresholded to 1 [resp. -1] when bigger than 1 [resp. lower than -1]) the magnetization of the configuration $\mathbf{s}$ along this mode. This allows us to define the probability

$$p_\ell[\mathbf{s}] = \frac{1 + m_\ell^\perp[\mathbf{s}]}{2}.$$

Then

$$s_\ell^\perp = 2\tau_\ell - 1, \qquad \forall\ell = 1,\dots\mathcal{N}[\mathbf{m}]$$

with τ_ℓ a Bernoulli variable of parameter $p_\ell[\mathbf{s}]$, gives us a set of spin variables fulfilling our needs.

Appendix D: Coulomb interaction picture

Let us consider the green function for the d-dimensional Laplacian ∇_d^2

$$K_d(|\mathbf{m}-\mathbf{m}'|) = \begin{cases}\frac{1}{2}|m-m'|, & (d=1)\\ \frac{1}{2\pi}\log|\mathbf{m}-\mathbf{m}'|, & (d=2)\\ -\frac{\Gamma\left(\frac{d}{2}-1\right)}{4\pi^{d/2}|\mathbf{m}-\mathbf{m}'|^{d-2}}, & (d>2)\end{cases}$$

which by definition is solution of

$$\nabla_d^2 K_d(|\mathbf{m}-\mathbf{m}'|) = \delta(\mathbf{m}-\mathbf{m}').$$

We can use it to rewrite $V[\mathbf{m}|\Theta]$ up to an irrelevant constant term as

$$V[\mathbf{m}|\Theta] = \int d^d\mathbf{m}'\rho(\mathbf{m}'|\Theta)K_d(|\mathbf{m}-\mathbf{m}'|) \tag{D1}$$

where ρ is the source term of the Poisson equation

$$\nabla_d^2 V[\mathbf{m}|\Theta] = \rho(\mathbf{m})$$

hence giving a density of Coulomb charges

$$\rho(\mathbf{m}|\Theta) = \sum_{j=1,\alpha=1}^{N_h,d} w_\alpha^2 v_j^{\alpha 2}\Big(1 - \tanh^2\big(\sqrt{N_v}\sum_{\beta=1}^{d} w_\beta m_\beta v_j^\beta - \theta_j\big)\Big).$$

To make sense of this quantity first remark that the function

$$\delta_\nu(x) \stackrel{\text{def}}{=} \frac{\nu}{2}\big[1-\tanh^2(\nu x)\big] \underset{\nu\to\infty}{\longrightarrow} \delta(x)$$

represents a normalized 1-d narrow density of width ν^{-1} such that ρ can be expressed as

$$\rho(\mathbf{m}|\Theta) = \frac{2}{N_v}\sum_{j=1}^{N_h}\nu_j\delta_{\nu_j}\big(\mathbf{n}_j^T\mathbf{m}-z_j\big) \tag{D2}$$

with ν_j and n_j given by equation (11,12). In this form, ρ is readily a superposition of N_h uniformly charged hyperplanes of finite width. Each hyperplane j being defined by a normal vector $\mathbf{n}_j$, an offset $z_j = \theta_j/\nu_j$ from the origin, a finite width ν_j^{-1} and a (hyper)surface charge density $2\nu_j/N_v$. Furthermore this can be decomposed into more elementary charged hyperplanes of zero width. Equation (D1) now rewrites

$$V[\mathbf{m}|\Theta] = \frac{2}{N_v}\sum_{j=1}^{N_h}\nu_j\int d^d\mathbf{m}'\delta_{\nu_j}\big(\mathbf{n}_j^T\mathbf{m}'-z_j\big)K_d(|\mathbf{m}-\mathbf{m}'|)$$

For each term j, writing $\mathbf{m}' = (\mathbf{n}_j^T\mathbf{m}')\mathbf{n}_j + \mathbf{m}'^\perp = z\mathbf{n}_j + \mathbf{m}'^\perp$, the transverse integration of $\mathbf{m}'^\perp$ yields

$$\int d\mathbf{m}' K_d\big(|\mathbf{m}-\mathbf{m}'|\big) =$$

$$\int dz d\mathbf{m}'^\perp K_d\big(\sqrt{(\mathbf{n}_j^T\mathbf{m}-z)^2 + (\mathbf{m}^\perp - \mathbf{m}'^\perp)^2}\big)$$

$$= \int dz|\mathbf{n}^T\mathbf{m}-z|.$$

up to an ill defined constant term after properly regularizing at large distances the integral over $\mathbf{m}'^\perp$. As a result the one particle potential takes the form

$$V[\mathbf{m}|\Theta] = \frac{2}{N_h}\sum_{j=1}^{N_h}\nu_j\int dz\delta_{\nu_j}\big(z-z_j\big)|\mathbf{n}_j^T\mathbf{m}-z|.$$

$$= \int d\mathbf{n}dz\; q(\mathbf{n},z)|\mathbf{n}^T\mathbf{m}-z|$$

with $q(\mathbf{n},z)$ given by equation (10)

Appendix E: Exact Coulomb charges interpolation

In order to interpolate exactly the empirical distribution $\hat{p}$ with a generalized Coulomb charges RBM based distribution it is needed to regularize $\log(\hat{p}(\mathbf{m}))$. This can be done in many different ways. Consider for instance

$$\delta_\epsilon(\mathbf{m}) \stackrel{\text{def}}{=} \frac{\exp\big(-\frac{|\mathbf{m}|^2}{2\epsilon}\big)}{(2\pi\epsilon)^{d/2}}$$

with infinitesimal ϵ to approximate our point-like distribution as

$$\hat{p}(\mathbf{m}) = \frac{1}{M}\sum_{k=1}^{M}\delta_\epsilon(\mathbf{m}-\mathbf{m}_k),$$

and let

$$q_\epsilon(k|\mathbf{m}) = \frac{\delta_\epsilon(\mathbf{m}-\mathbf{m}_k)}{\sum_l \delta_\epsilon(\mathbf{m}-\mathbf{m}_l)}.$$

These probability weights realize a smooth partition of the space at finite ϵ with Voronoi cells $\mathcal{R}_k$ centered at each data point, $q_\epsilon(k|\mathbf{m})$ representing the probability that $\mathbf{m}$ belongs to kth cell. Equipped with this notation we have

$$\nabla_\mathbf{m}\delta_\epsilon(\mathbf{m}-\mathbf{m}_k) = -\frac{\mathbf{m}-\mathbf{m}_k}{\epsilon}\delta_\epsilon(\mathbf{m}-\mathbf{m}_k)$$

$$\nabla_\mathbf{m} q_\epsilon(k|\mathbf{m}) = -\frac{\mathbf{m}-\mathbf{m}_k}{\epsilon}q_\epsilon(k|\mathbf{m}) + \sum_{\ell=1}^{M}\frac{\mathbf{m}-\mathbf{m}_\ell}{\epsilon}q_\epsilon(k|\mathbf{m})q_\epsilon(\ell|\mathbf{m}).$$

As a result we get

$$\nabla_d^2\log\hat{p}(\mathbf{m}) = -\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\sum_{k=1}^{M}\mathrm{Var}_{k\sim q_\epsilon(k|\mathbf{m})}[\mathbf{m}_k].$$

When ϵ becomes small compared to nearest neighbour distances this quantity becomes constant ($= -1/\epsilon$) except on the intersections between Voronoi cells, in particular on common faces $\mathcal{R}_k\cap\mathcal{R}_\ell$ between two cells $\mathcal{R}_k$ and $\mathcal{R}_\ell$ it is

$$\nabla_d^2\log\hat{p}(\mathbf{m}) \underset{\epsilon\to 0}{\sim} -\frac{1}{\epsilon} + \frac{|\mathbf{m}_k-\mathbf{m}_\ell|}{2\epsilon}\delta(\mathbf{m}\in\mathcal{R}_k\cap\mathcal{R}_\ell).$$

Indeed, let

$$\theta_k \stackrel{\text{def}}{=} \frac{1}{2}(\mathbf{m}_k + \mathbf{m}_{k+1}) \qquad \text{and} \qquad \Delta_k \stackrel{\text{def}}{=} \frac{1}{2}(\mathbf{m}_{k+1} - \mathbf{m}_k).$$

For $\delta\mathbf{m} = \mathbf{m} - \theta_k$ small compared to Δ_k we have

$$q_k(\theta_k + \delta\mathbf{m}) = \frac{1}{2}\Big[1 - \tanh\Big(\frac{\Delta_k^T \delta\mathbf{m}}{2\epsilon}\Big)\Big],$$

leading to

$$\mathrm{Var}_{k\sim q_k(\mathbf{m})}[\mathbf{m}_k] = |\Delta_k|^2\Big[1 - \tanh^2\Big(\frac{\Delta_k^T \delta\mathbf{m}}{2\epsilon}\Big)\Big].$$

Since $\frac{\nu}{2}\big[1 - \tanh^2(\nu x)\big]$ tends to $\delta(x)$ when $\nu \to \infty$ we arrive at the statement. As a result the distribution of charges is composed of a constant background + surface distribution on Voronoi cells intersections:

$$\rho_{\text{bulk}}(\mathbf{m}) = -\frac{1}{N_v\epsilon} + \frac{|\mathbf{m}_k - \mathbf{m}_\ell|}{2N_v\epsilon}\delta(\mathbf{m} \in \mathcal{R}_k \cap \mathcal{R}_\ell).$$

The Voronoi cells intersecting with the boundary of the $\mathbf{m}$ domain induce additional surface charges which can be directly taken care of with visible bias. Let us show how this works in 1-d. Let us call

$$V(m) = \frac{1}{N_v}\log\hat{p}(m)$$

which when regularized reads

$$V(m) = -\frac{1}{N_v}\min_k \frac{(m - m_k)^2}{2\epsilon}.$$

From what precedes, this potential can be exactly decomposed onto a set of features as

$$V(m) = -\frac{m^2}{2N_v\epsilon} + \eta m + \sum_j q_j|m - z_j|$$

with

$$q_j = \frac{1}{2N_v\epsilon}(m_{j+1} - m_j),$$

while from the limit behaviour $V'(1)$ and $V'(-1)$ of $V'(m)$ we get

$$\eta = \frac{m_1 + m_{N_v}}{2\epsilon}.$$

Appendix F: RBM optimization seen as a linear regression

The projection of the empirical distribution onto the space of RBM with finite number of features is classically done by minimizing the Kullback Liebler divergence (D_{KL}). If however our RBM space is chosen with a high number of relevant features, we may expect the solution to be close enough to the empirical distribution so that a Fisher metric, i.e. the infinitesimal counterpart of the D_{KL}, evaluated from the solution or from the empirical distribution should coincide. In that case it might be pertinent to use it instead of the D_{KL}. Let us formalize more precisely this projection problem. On one hand we have the empirical measure approximated by a Coulomb based RBM model of the form

$$p(\mathbf{m}|\hat{\rho}) = \frac{1}{Z[\hat{\rho}]}e^{-N_v\mathcal{F}(\mathbf{m}|\hat{\rho})}$$

with

$$\mathcal{F}(\mathbf{m}|\hat{\rho}) = \mathcal{F}^\perp[\mathbf{m}] - \mathcal{S}[\mathbf{m}] - \int d\mathbf{m}'\hat{\rho}(\mathbf{m}')K_d(|\mathbf{m} - \mathbf{m}'|),$$

where $\hat{\rho}(\mathbf{m})$ is, as seen in the previous Section, the charge density concentrated on the Voronoi cells faces coming from the empirical part $\log\hat{p}(\mathbf{m})$ (including surface terms at the edge of the domain of $\mathbf{m}$). On the other hand we have an RBM with a pointwise distribution of features $q(\mathbf{n}, z)$ yielding a free energy of the form

$$\mathcal{F}(\mathbf{m}|\Theta) = \mathcal{F}^\perp[\mathbf{m}] - \mathcal{S}[\mathbf{m}] - \int d\mathbf{m}'\rho(\mathbf{m}'|\Theta)K_d(|\mathbf{m} - \mathbf{m}'|),$$

with

$$\rho(\mathbf{m}|\Theta) = \sum_{j=1}^{N_h} q_j\delta\big(\mathbf{n}_j^T\mathbf{m} - z_j\big).$$

Our goal is to find the (positive) weights $\{q_j, j = 1, \dots N_h\}$ such that the following distance

$$D(\rho, \rho') = \int d\mathbf{m}_1 d\mathbf{m}_2 \rho(\mathbf{m}_1)J(\mathbf{m}_1, \mathbf{m}_2)\rho'(\mathbf{m}_2), \quad \text{(F1)}$$

between $\hat{\rho}$ and ρ is minimized. Here the relevant metric J is the Fisher metric defined as

$$\begin{aligned} J\big[\mathbf{m}_1, \mathbf{m}_2\big] &= \mathrm{Cov}_{\mathbf{m}\sim p(\mathbf{m}|\Theta)}\Big[K_d(|\mathbf{m} - \mathbf{m}_1|), K_d(|\mathbf{m} - \mathbf{m}_2|)\Big], \\ &\simeq \mathrm{Cov}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}\Big[K_d(|\mathbf{m} - \mathbf{m}_1|), K_d(|\mathbf{m} - \mathbf{m}_2|)\Big], \end{aligned} \quad \text{(F2)}$$

approximated at the empirical point in last equation. As we shall see this projection turns out to be a linear regression of the centered random variable

$$\begin{aligned} V(\mathbf{m}|\hat{\rho}) = &\int d\mathbf{m}'\hat{\rho}(\mathbf{m}')K_d(|\mathbf{m} - \mathbf{m}'|) \\ &- \mathbb{E}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}\Big[\int d\mathbf{m}'\hat{\rho}(\mathbf{m}')K_d(|\mathbf{m} - \mathbf{m}'|)\Big] \end{aligned}$$

onto the set of centered random variables (the score variables associated with $\boldsymbol{q}$)

$$V_j(\mathbf{m}) \stackrel{\text{def}}{=} \int d\mathbf{m}'\delta\big(\mathbf{n}_j^T\mathbf{m}' - z_j\big)K_d(|\mathbf{m}-\mathbf{m}'|)$$
$$- \mathbb{E}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}\Big[\int d\mathbf{m}'\delta\big(\mathbf{n}_j^T\mathbf{m}' - z_j\big)K_d(|\mathbf{m}-\mathbf{m}'|)\Big]$$
$$= |\mathbf{n}_j^T\mathbf{m} - z_j| - \mathbb{E}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}\Big[|\mathbf{n}_j^T\mathbf{m} - z_j|\Big].$$

$\mathbb{E}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}$ and $\mathrm{Cov}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}$ denote respectively empirical expectation and covariance, according to our assumption that the solution is close to $\hat{p}$. Indeed, from elementary linear algebra, the orthogonal projection $V^{\parallel}$ of a given vector $\hat{V}$, onto a subspace spanned by a set of independent vectors V_k is given by

$$V^{\parallel} = \sum_{k,l}\Big[G^{-1}\Big]_{kl}(V_l,\hat{V})V_k \tag{F3}$$

with $G_{kl} = (V_k, V_l)$ the Gram matrix of the set of vector V_k for some given inner product $(\cdot,\cdot)$. Specified to our problem, the vectors are the densities ρ, or equivalently the random variables $V(\mathbf{m}|\rho)$ with inner product (F1,F2) resulting in

$$\big(V(\mathbf{m}|\rho), V(\mathbf{m}|\rho')\big) = \mathrm{Cov}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}\Big[V(\mathbf{m}|\rho), V(\mathbf{m}|\rho')\Big].$$

The projection of $\hat{V}[\mathbf{m}] = V(\mathbf{m}|\hat{\rho})$ is then given by $V^{\parallel}$ in (F3) with G the empirical covariance matrix of $\{V_1, \dots V_{N_h}\}$ and $(V_k, \hat{V})$ the empirical covariance between V_k and $\hat{V}$ (if the set V_k is not independent the pseudo-inverse of G is taken instead of G^{-1}). At this point this regression seems intractable since $\hat{V}$ involves a very complicated density of charge $\hat{\rho}$. This is not the case because by construction we have

$$\int d\mathbf{m}'\hat{\rho}(\mathbf{m}')K_d\big(|\mathbf{m}'-\mathbf{m}|\big) = \frac{1}{N_v}\log\hat{p}(\mathbf{m}) - \mathcal{S}[\mathbf{m}] + \mathcal{F}^{\perp}[\mathbf{m}],$$

and $\log\hat{p}(\mathbf{m}) = \log\frac{1}{M}$ when evaluated on the data. This means that the solution to our projection problem is obtained by performing the previous linear regression with

$$\hat{V}(\mathbf{m}) = \mathcal{F}^{\perp}[\mathbf{m}] - \mathcal{S}[\mathbf{m}] - \mathbb{E}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}\Big[\mathcal{F}^{\perp}[\mathbf{m}] - \mathcal{S}[\mathbf{m}]\Big],$$

so that the RBM distribution will be finally of the form

$$P_{\mathrm{RBM}}(\mathbf{m}) = \frac{1}{Z}e^{-N_v\mathcal{F}(\mathbf{m}|\Theta)}$$

with

$$\mathcal{F}(\mathbf{m}|\Theta) = \mathcal{F}^{\perp}[\mathbf{m}] - \mathcal{S}[\mathbf{m}] - \sum_{j=1}^{N_h} q_j|\mathbf{n}_j^T\mathbf{m} - z_j|,$$
$$\stackrel{\text{def}}{=} V(\mathbf{m}) - V_{\mathrm{RBM}}(\mathbf{m}),$$

i.e. a difference between 2 convex potential whenever $\mathcal{F}^{\perp}[\mathbf{m}]$ is convex or negligible. The interpolation point corresponding to the situation where there is a sufficient amount of Coulomb features to model exactly the empirical distribution is shown on Fig. 4. In this appealing picture there is however a shortcoming. The fact that we impose the features weights to be non-negative insures the regression curve to be convex but do not prevent it to pass above the fitted potential in empty regions of data. The reason for this, while the information theory argument provides us with a strong guarantee at first sight, is that the empirical Fisher metric is not relevant everywhere on the embedding functional space defined by the RBM features used to approximate $V(\mathbf{m})$, but only on regions supported with data. Other directions are represented by random variables which are decorrelated from the data, so the Fisher metric is not covered by (F1) along these directions. This requires the linear regression to be complemented with some additional regularization which as shown on Fig. 4 should involve also the distance between the derivatives of the two profiles measured at the sample points.

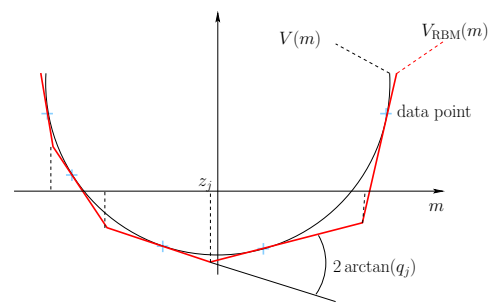

FIG. 4. Picture of the $V_{\mathrm{RBM}}(\mathbf{m})$ (in red) at the interpolation threshold in 1-d.

Appendix G: Experiments

The transverse free energy being absent and the entropy term independent of $\Theta \equiv \boldsymbol{q}$, the loss optimized in the experiments is given by

$$\mathcal{L}[\boldsymbol{q}] = -\mathbb{E}_{\mathbf{m}\sim\hat{p}}\big[V[\mathbf{m}|\boldsymbol{q}]\big] - \log\big(Z[\boldsymbol{q}]\big),$$

corresponding to

$$p(\mathbf{m}|\boldsymbol{q}) = \frac{1}{Z[\boldsymbol{q}]}e^{-N_v\mathcal{F}[\mathbf{m}|\boldsymbol{q}]}$$

with

$$\mathcal{F}[\mathbf{m}|\boldsymbol{q}] = -\mathcal{S}[\mathbf{m}] - \sum_{j=0}^{N_h} q_j|\mathbf{n}_j^T\mathbf{m} - z_j|$$

In the 1-d study case, we have $n_j = 1$ and the discretiza-

TABLE I. LL obtained for the RBM in the study cases

	LL_{Ref}	LL_{Coulomb}	LL_{RBM}	N_{features}	N_{epochs}	N^d_{pts}	γ
$d=1$	-471.45	-479.85	-479.82	5	5000	100	0.0001
	-471.45	-471.66	-471.69	10	7000	100	0.0001
	-471.45	-471.48	-471.48	20	10000	100	0.001
	-471.45	—	-535.00	20	2500	—	0.001
$d=2$	-621.90	-623.74	-623.61	49	113000	900	0.0001
	-621.90	-622.16	-622.24	169	155000	900	0.0001
	-621.90	-621.97	-631.79	900	410000	1600	0.0001

tion concerns only $z_j = -1 + 2j/N_h, j = 1, \dots N_h$. The features $j = 0$ and $j = N_h$ correspond to the visible bias η.

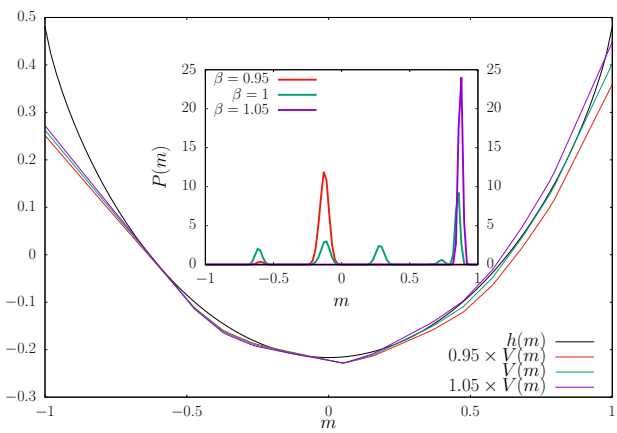

FIG. 5. First order phase transition mechanism of the "Coulomb" RBM illustrated on the 1-d case. β represents an annealing inverse temperature, which effect is to multiply the energy part $V(m)$ in the free energy $\mathcal{F}_\beta(m) = \beta V(m) - h(m)$. $\beta = 1$ is a reference RBM where many states are present with comparable probabilistic weights; for $\beta > 1$ [resp. $\beta < 1$] the state with highest [resp. lowest] $V(m)$ is favored.

In the 2-d study case we have $\mathbf{n}_j = \big(\cos(p\pi/\sqrt{N_h}), \sin(p\pi/\sqrt{N_h})\big)$ and $z_j = -1 + 2r/\sqrt{N_h}$ with the entire decomposition of $j = p\sqrt{N_h} + r$. The continuous dynamics of the parameters corresponding to an ordinary gradient reads

$$\dot{\boldsymbol{q}} = \gamma \nabla_{\boldsymbol{q}} \mathcal{L}[\boldsymbol{q}],$$

with the learning rate γ. The scores associated with these parameters are centered random variables classically defined as

$$\boldsymbol{q}^*(\mathbf{m}) = \nabla_{\boldsymbol{q}} \log p(\mathbf{m}|\boldsymbol{q}),$$

which in our case read:

$$q^*_j(\mathbf{m}) = |\mathbf{n}_j^T \mathbf{m} - z_j| - \mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}|\boldsymbol{q})}\big[|\mathbf{n}_j^T \mathbf{m} - z_j|\big].$$

In terms of these variables the gradient of the log likelihood simply reads

$$\nabla_{\boldsymbol{q}} \mathcal{L}[\boldsymbol{q}] = \mathbb{E}_{\mathbf{m}\sim\hat{p}}\big[\boldsymbol{q}^*(\mathbf{m})\big],$$

$\hat{p}$ being the empirical data distribution. In a continuous limit of the training process indexed by the training time t, let $\hat{\boldsymbol{q}}^*_t$ the time dependent expectation of $\boldsymbol{q}_t(\mathbf{m})$ w.r.t the empirical distribution $\hat{p}$. We have

$$\hat{q}^*_{j,t} = \int d\mathbf{m}\big(\hat{p}(\mathbf{m}) - p(\mathbf{m}|\boldsymbol{q}_t)\big)|\mathbf{n}_j^T \mathbf{m} - z_j|.$$

The evolution of $\hat{\boldsymbol{q}}^*_t$ with time is given by

$$\frac{d}{dt}\hat{\boldsymbol{q}}^*_t = -\gamma \mathrm{Cov}_{\mathbf{m}\sim p(\mathbf{m}|\boldsymbol{q})}\big[\boldsymbol{q}^*_t(\mathbf{m}), \boldsymbol{q}^{*T}_t(\mathbf{m})\big]\hat{\boldsymbol{q}}^*_t,$$

where $\mathrm{Cov}_{\mathbf{m}\sim p(\mathbf{m}|\boldsymbol{q})}$ denotes the covariance under $p(\mathbf{m}|\boldsymbol{q})$ and corresponds to the Fisher metric of the $\boldsymbol{q}$ parameter space. As we see as long as the covariance matrix is strictly positive definite the dynamics is contractant. When using the natural gradient [32] defined here as

$$\tilde{\nabla}_{\boldsymbol{q}} = \mathrm{Cov}_{\mathbf{m}\sim p(\mathbf{m}|\boldsymbol{q}_t)}\big[\boldsymbol{q}^*_t(\mathbf{m}), \boldsymbol{q}^{*T}_t(\mathbf{m})\big]^{-1}\nabla_{\boldsymbol{q}},$$

the dynamics simplifies to

$$\frac{d}{dt}\hat{\boldsymbol{q}}^*_t = -\gamma \hat{\boldsymbol{q}}^*_t.$$

The norm of $\hat{\boldsymbol{q}}^*_t$ can be monitored during learning allowing for an adaptive strategy for the learning rate used in practice in these experiments to control the convergence as well as the stopping criterion. Thanks to the choice made for u_α in these experiments the number of configurations $\mathbf{m}$ is equal to $N+1$ and $(N+1)^2/4$ respectively in the 1 and 2-d case, so the LL of the resulting "Coulomb" RBM which are obtained can be evaluated exactly in both cases as well as the corresponding standard RBM. The latter is obtained through the following asymptotic mapping assuming that $\sqrt{\sum_i W_{ij}^2}$ is large for all j:

$$\log\cosh\Big(\sum_{i=1}^{N_v} W_{ij} s_i - \theta_j\Big) \approx \sqrt{N_v \sum_i W_{ij}^2}\Big|\frac{\mathbf{s}^T}{\sqrt{N_v}}\mathbf{n}_j - z_j\Big|$$

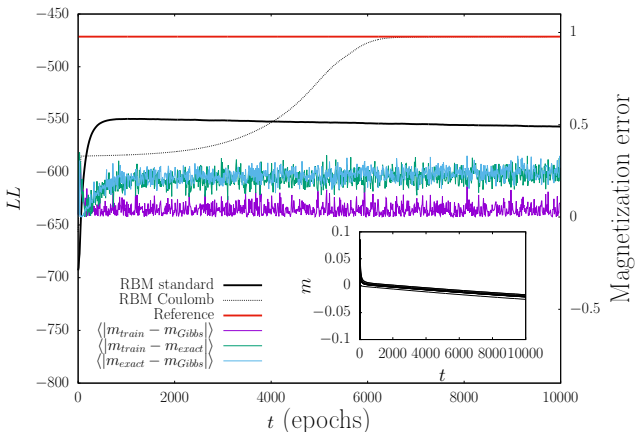

FIG. 6. Illustration of the failure of the standard RBM training due the occurrence of a first order transition signaled here by a divergence between the Gibbs sampling evaluation of the magnetization along the intrinsic axis with the exact one (main plot), and due to the trapping of the z_j's close to zero (inset).

with

$$n_{ij} = \frac{W_{ij}}{\sqrt{\sum_i W_{ij}^2}} \qquad \text{and} \qquad z_j = \frac{\theta_j}{\sqrt{N_v \sum_i W_{ij}^2}}$$

leading to

$$q_j = \sqrt{\frac{1}{N_v} \sum_i W_{ij}^2}.$$

This results into the following correspondence

$$W_{ij} = \sqrt{N_v} q_j n_{ij}, \qquad \theta_j = N_v q_j z_j.$$

In both experiments the number of visible variables is $N_v = 10^3$, and the LL is an average value estimated on an independent test set of 10^3 samples. The number of points N^d_{pts} used to estimate the integrals over $\mathbf{m}$ needed to compute the natural gradient give a contribution $N_f N^d_{\text{pts}}$ to the complexity in a naive setting. The indicated values in the Table I correspond to the point where the results become insensitive to N^d_{pts}. The values LL_{Coulomb} and LL_{RBM} measured respectively for the "Coulomb" machine which is optimized with the natural gradient and its corresponding RBM using the previous mapping are reported on Table I are compared with the reference value LL_{Ref} of the hidden mixture model used to generate the data. Note that the mapping gives a poor model when many weak features are used as in the $d = 2$ case with $N_h = 900$. Note also that using the ordinary gradient instead of the natural one seems to keep the last bits of LL out of reach in a reasonable time.

Finally Fig. 6 illustrates the reasons for the failure of the standard RBM training. First the Gibbs sampling procedure is plagued by the presence of 1st order phase transitions which is well understood when considering the "Coulomb" RBM. Indeed, in that case changing the temperature corresponds to multiplying all the feature weights q_j by a common factor β representing for instance an annealing inverse temperature. The learning procedure is supposed to tune precisely the difference $\mathcal{F}_\beta[\mathbf{m}|\boldsymbol{q}] = \beta V(\mathbf{m}|\boldsymbol{q}) - \mathcal{S}[\mathbf{m}]$ at $\beta = 1$, in order to obtain many coexisting states corresponding to different values of condensed magnetization $\mathbf{m}$. Then as in the example of Fig. 5, changing slightly β has the effect of concentrating the probability distribution on the state with highest or lowest value of $V(\mathbf{m}|\boldsymbol{q})$ depending on whether β is smaller or greater than one. The second source of failure is, as expected from the electrostatic picture, that the hidden bias given in rescaled form by $z_j = \theta_j/\nu_j$ along with (11) get trapped, around zero in the example of Fig. 6 which prevent the machine to form more than two ferromagnetic states in 1d.