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 on it by considering the situation where the data have low intrinsic dimension, offering the possibility of an exact treatment, illustrated with simple study cases where the standard training is observed to fail. The reason for this failure is clarified thanks to a Coulomb interactions reformulation of the model where the hidden features are uniformly charged hyperplanes interacting repulsively with point-like charges representing the data. This leads us to consider a convex relaxation of the original optimization problem thereby resulting in a unique solution, obtained in precise numerical form in $d = 1, 2$ while a constrained linear regression solution is 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 use 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] which remains very popular thanks to its simplicity and effectiveness when applied to moderately high dimensional data [11–13]. It is a 2-layers undirected neural network which represents the data in the form of a Gibbs distribution of visible and latent variables (see Figure 1):

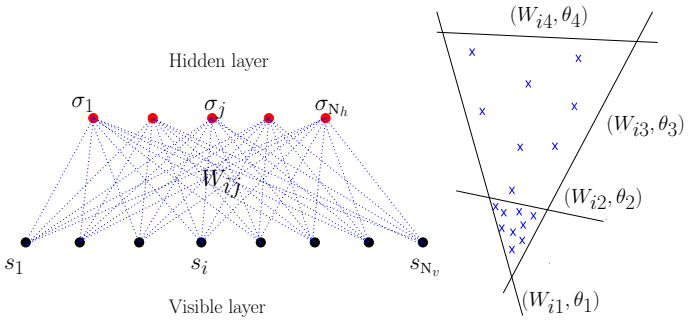

FIG. 1. Bipartite structure of the RBM (left). Hyperplanes defined by the weight vectors and bias associated to each hidden variable can delimit fixed density regions in input space (right).

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

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 to 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 Figure 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

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

This is a non-trivial optimization problem in two respects: it is non convex and the loss function $-Ł[\Theta]$ is difficult to estimate because $\log\big(Z[\Theta]\big)$ is not tractable. Nevertheless, the gradient $\nabla_\Theta Ł[\Theta]$ 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 as been found in [21], retrieval capacity has been characterized in [22, 23] and compositional mechanisms analyzed in [24, 25] (see more recent ref. 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 where a certain number of modes emerge from a Marchenko-Pastur bulk at initialization and condense to built up a structured ferromagnetic phase. In this letter we complete this program by showing that the two main difficulties (non-tractability and non-convexity) of the training can be addressed at least in the special case where the number of condensed modes is small i.e. when training data lay on a flat intrinsic space of low dimension.

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 through its SVD decomposition as

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

with w_α, $\boldsymbol{u}^\alpha$ and $\boldsymbol{v}^\alpha$ representing respectively the singular values, the 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.

$$m_\alpha \stackrel{\text{def}}{=} \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} s_i u_i^\alpha = \mathcal{O}(1).$$

For an RBM trained on some data, d would represent their intrinsic dimension at least locally. 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 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 given 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 $\{s_1^\perp, \dots, s_{\mathcal{N}[\mathbf{m}]}^\perp\}$. In terms of this new set of visible variables we may now formally write our distribution as

$$P[\mathbf{m}, \mathbf{s}^\perp|\Theta] = \frac{e^{-N_v \mathcal{F}^\parallel[\mathbf{m}|\Theta] - \mathcal{H}_{\text{eff}}\left[\mathbf{s}^\perp|\mathbf{m},\Theta\right]}}{\int d\mathbf{m}\ e^{-N_v \mathcal{F}[\mathbf{m}|\Theta]}},$$

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:

$$\mathcal{F}^\parallel[\mathbf{m}|\Theta] = -\mathcal{S}[\mathbf{m}] - \sum_{\alpha=1}^{d} \eta_\alpha m_\alpha - V[\mathbf{m}|\Theta], \tag{5}$$

$$\mathcal{F}^\perp[\mathbf{m}|\Theta] = -\frac{1}{N_v} \log\Big(\frac{1}{2^{\mathcal{N}[\mathbf{m}]}} \sum_{\mathbf{s}^\perp} e^{-\mathcal{H}_{\text{eff}}[\mathbf{s}^\perp|\mathbf{m},\Theta]}\Big), \tag{6}$$

($\eta_\alpha \stackrel{\text{def}}{=} \frac{1}{\sqrt{N_v}} \sum_i \eta_i u_i^\alpha$) to which are respectively associated a potential function for the magnetizations and the Hamiltonian for the transverse degrees of freedom:

$$V[\mathbf{m}|\Theta] = \frac{1}{N_v} \sum_{j=1}^{N_h} \log\cosh\big(\sqrt{N_v} \sum_{\alpha=1}^{d} w_\alpha m_\alpha v_j^\alpha - \theta_j\big), \tag{7}$$

$$\mathcal{H}_{\text{eff}}\left[\mathbf{s}^\perp|\mathbf{m},\Theta\right] = \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} \eta_\ell^\perp[\mathbf{m},\Theta] s_\ell^\perp + \sum_{\ell,\ell'=1}^{\mathcal{N}[\mathbf{m}]} W_{\ell\ell'}^\perp[\mathbf{m},\Theta] s_\ell^\perp s_{\ell'}^\perp$$

By convenience we assign to $\mathcal{F}^\parallel$ the default entropy ($\mathcal{N}[\mathbf{m}]\log(2)$) contribution of the transverse variables in order that $\mathcal{F}^\perp$ vanishes when $\mathcal{H}_{\text{eff}} = 0$. The couplings $W_{\ell\ell'}^\perp[\mathbf{m},\Theta]$, local fields $\eta_\ell^\perp[\mathbf{m},\Theta]$ and consistent definitions of transverse variables $s_\ell^\perp$ are given in the Appendix B. This defines for each $\mathbf{m}$ a disordered Ising model of $\mathcal{N}[\mathbf{m}]$ spins with paramagnetic-like state of order.

Coulomb formulation and linear regression. The potential term in $\mathcal{F}^\parallel$ which act on the magnetization $\mathbf{m}$ representing here the position of a particle in a d-dimensional space can be written as (See appendix C)

$$V[\mathbf{m}|\Theta] = \int d\mathbf{n} d\theta\ \Theta(\mathbf{n},\theta) |\mathbf{n}^T\mathbf{m} - \theta|, \tag{8}$$

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

$$\Theta(\mathbf{n},\theta) = \frac{2}{N_v} \sum_{j=1}^{N_h} \nu_j \delta_{\nu_j}\Big(\frac{\theta_j}{\nu_j}\Big) \delta(\mathbf{n} - \mathbf{n}_j) \delta\Big(\theta - \frac{\theta_j}{\nu_j}\Big). \tag{9}$$

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

$$\nu_j = \sqrt{N_v}\Big(\sum_{\alpha=1}^{d} w_\alpha^2 v_j^{\alpha\,2}\Big)^{\frac{1}{2}}, \tag{10}$$

$$n_j^\alpha = \frac{\sqrt{N_v}}{\nu_j} w_\alpha v_j^\alpha. \tag{11}$$

The kernel $|\mathbf{n}_j^T\mathbf{m}-\theta|$ represents the Coulomb potential exerted by a uniformly charged hyperplane defined by its normal vector $\mathbf{n}$ and its distance θ to the origin to a charge located at $\mathbf{m}$. As a result, each feature j corresponds also to charged hyperplane of normal vector $\mathbf{n}_j$, offset θ_j/ν_j but of finite width of order ν_j^{-1}. At this point let us remark that the singular values w_α control two different things through ν_j, namely the strength of the Coulomb interaction via (8,9,10) and the width of the charged hyperplanes while the right singular vectors projection on the various modes v_j^α control the orientation of these hyperplanes in the intrinsic space through (11). Overall the density of Coulomb charges in the d-dimensional intrinsic space is given by

$$\rho(\mathbf{m}) = \int d\mathbf{n}d\theta\ \Theta(\mathbf{n},\theta)\delta\big(\mathbf{n}^T\mathbf{m}-\theta\big). \tag{12}$$

If not constrained to be a superposition of a finite number of charged hyperplanes, ρ can be adjusted to have the following matching with any distribution $\hat{p}(\mathbf{m})$:

$$e^{N_v\left(\mathcal{S}(\mathbf{m})-\mathcal{F}^\perp[\mathbf{m}|\rho]+\int d\mathbf{m}'\rho(\mathbf{m}')K_d(|\mathbf{m}-\mathbf{m}'|)\right)} \propto \hat{p}(\mathbf{m}),$$

with $K_d(|\mathbf{m}-\mathbf{m}'|)$ the inverse of the d-dimensional Laplacian Δ_d. Inverting this yields

$$\rho(\mathbf{m}) = \Delta_d\big(\frac{1}{N_v}\log\hat{p}(\mathbf{m}) - \mathcal{S}[\mathbf{m}] + \mathcal{F}^\perp[\mathbf{m}|\rho]\big) \tag{13}$$

up to surface terms. As long as $\mathcal{F}^\perp$ is independent of ρ this constitutes an explicit solution to the problem which has to be approximated in the form of (12). 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 approximators [30]. 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 Θ. The log likelihood of the RBM has then three terms

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

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

$$\mathbb{E}_{\hat{p}}\big[V[\mathbf{m}|\Theta]\big] = \int d\mathbf{m}d\mathbf{n}d\theta\ \hat{p}(\mathbf{m})|\mathbf{n}_j^T\mathbf{m}-\theta|\Theta(\mathbf{n},\theta), \tag{14}$$

takes the form of a repulsive Coulomb interaction between data points and charged hyperplanes. The optimization of $Ł[\Theta]$ w.r.t. the features weights $\Theta(\mathbf{n},\theta)$ instead of the original RBM parameters is convex when $\mathcal{F}^\perp$ is independent of Θ, as this "Coulomb" formulation is in the exponential family. It corresponds to a slight extension of the RBM model in terms of more general activation function (including RELU [31] for instance and similar also to [32]), where each feature contribution in (3) comes with a non-negative weight Θ_j which from now on is the parameter to be optimized, while the features themselves defined by the pairs $(\mathbf{n}_j,\theta_j)$ are predefined. Solving equation (13) with a smoothing of $\log\hat{p}(\mathbf{m})$ leads

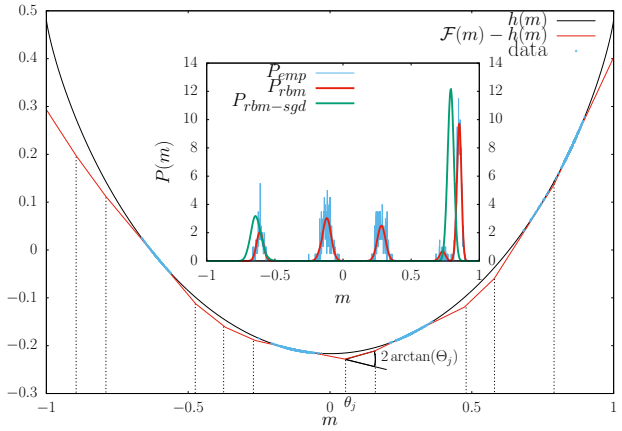

FIG. 2. 1-d intrinsic data ($N_v = 10^3$) with 5 clusters solved with $N_h = 20$ predefined features thanks to a natural gradient ascent of the LL. Dotted lines indicate location of features with non-vanishing weights Θ_j. The feature contributions $\mathcal{F}(m)-h(m)$ to the free energy is seen to regress $h(m)$ on the data. The resulting distribution is shown (red) on the inset with the empirical training distribution (blue) and the result of a standard RBM training (green).

to overfit the data with a density of Coulomb charges concentrated on the faces of the Voronoi cells enclosing the data points (see Appendix D). This should be projected on a density ρ of the form (12) corresponding to a finite number of features to be meaningful. This projection to be consistent need to be done using the Fisher metric [33] instead of the Euclidean one for instance and end 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 LL. Suppose now we expect the optimal solution to be very close to $\hat{p}$. This allows us to consider instead the Fisher metric estimated at the empirical point $\hat{p}$ and turn the problem into the following linear regression

$$\Theta^\star = \underset{\Theta}{\mathrm{argmin}}\ \mathbb{E}_{\hat{p}}\Big[\big|\mathcal{S}(\mathbf{m}) + \sum_{j=1}^{N_h}\Theta_j V_j(\mathbf{m})\big|^2\Big]$$

of $\mathcal{S}[\mathbf{m}]$ on the score variables $V_i(\mathbf{m}) \overset{\mathrm{def}}{=} \frac{\partial\log(p(\mathbf{m}|\Theta))}{\partial\Theta_j}$ conjugate to Θ_j (see Appendix E).

Study case. 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. We assume a rank one weight matrix of the form $W = w_1 u^1 v^{1T}$ since we expect the singular value associated to other directions 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

$$\mathcal{F}[m] = h(m) - \sum_{j=1}^{N_v} \Theta_j |m - \theta_j|, \qquad (\Theta_j \geq 0)$$

with $h(m) = \frac{1}{2}(1 \pm m)\log(1 \pm m)$. Performing the natural gradient ascent [33] of the LL yield optimal solution as the one shown on Figure 2. Note that for this kind of problem, to the best of our training experiments with any type of sampling and extensive meta-parameters settings, the standard RBM training basically fails to resolve properly the cluster structure, essentially because the charged hyperplanes remain trapped into sub-optimal configurations related to the Coulomb barriers formed by the clusters of data. In the convex "Coulomb" setting this task becomes much easier. As is manifest on Figure 2 the feature part of the free energy ($\mathcal{F}(m) - h(m)$) is performing a linear regression of $h(m)$ in terms of a piece-wise linear function where the break points corresponds to the locations θ_j of the relevant features and Θ_k the corresponding break of slope at these points. This regression involves however an implicit regularization 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 ceases to be a meaningful approximation to the D_{KL}. Finding this regularization remains to us an open question which if identified could lead to an approach similar in spirit to support vector machines [34] for classification problems. As it appears, the linear parts are nearly tangent to $h(m)$ at the data points and intersect at the position of relevant features. This suggests that a relevant set of candidate features could possibly be obtained more generally by intersecting tangent hyperplanes to the hypersurface $S(\mathbf{m}) = cte$ in $\mathbb{R}^{N_v+1}$ at the datapoints.

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 such that the weight matrix once the relevant axes have been found, is of the form $W = w_1 u^1 v^{1T} + w_2 u^2 v^{2T}$. We have then a finite magnetization (m_1, m_2) along each direction and the free energy considered with a continuous field of features in the Coulomb formulation reads

$$\mathcal{F}[\mathbf{m}|\Theta] = \frac{1}{2}[h(m^+) + h(m^-)] - \int d\omega d\theta \; \Theta(\omega, \theta)|m_1 \cos(\omega) + m_2 \sin(\omega) - \theta|,$$

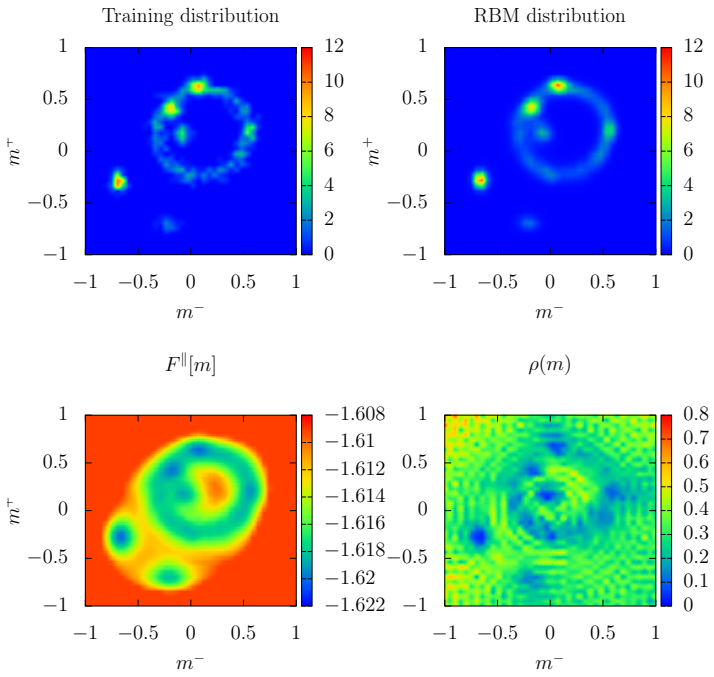

FIG. 3. Free energy landscape (bottom left) found with $N_h = 900$ pre-defined features on a 2-d intrinsic dataset ($N_v = 10^3$) with 6 point-like clusters and a circular one and corresponding Coulomb charges distribution obtained here (bottom right)

where $m^{\pm} = m_1 \pm m_2$ and $\omega \in [0, \pi[$ is the angle made by the normal vector $\mathbf{n}$ to the charged lines with the m_1 axe. Again this can be optimized efficiently by performing the natural gradient ascent of the LL as is shown on Figure 3. Here a large number of features have been predefined on a regular lattice of the (ω, θ) plane, in order to obtain a continuous charge distribution and a smooth free energy landscape. Again, for this kind of problem our experiments (not shown) indicate that standard RBM training procedures fail by a large margin.

Conclusion. The specific treatment presented here could be used directly as it is for practical applications involving intrinsically 3-d data but presumably not beyond $d = 3$. Nevertheless our main goal was to understand better the optimization problem posed by the RBM training. The Coulomb relaxation presented here opens many options for algorithmic developments, in order in particular to address the shortcomings of the RBM model identified in this letter and to explore further the 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] N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.

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

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

[33] S-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[34] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152. ACM, 1992.

[35] H. Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478:1–69, 2009.

Appendix A: Canonical ensemble with 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

$$m_\alpha = \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} s_i u_i^\alpha \stackrel{\text{def}}{=} s_\alpha, \qquad \alpha = 1, \dots d,$$

of variables in $[-1,1]$ and $\mathcal{N}[\mathbf{m}]$ transverse spin variables. The change of measure is made by looking at the prior distribution over the original spin variables:

$$f_0[\mathbf{s}] = \frac{1}{2^{N_v}} = f[\mathbf{s}^\perp|\mathbf{m}]f[\mathbf{m}],$$

where

$$f[\mathbf{m}] = \frac{1}{2^{N_v}} \sum_{\mathbf{s}} \prod_{\alpha=1}^{d} \delta(s_\alpha - m_\alpha) = e^{N_v(\mathcal{S}[\mathbf{m}]-\log 2)} \quad \text{(A1)}$$

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

$$f[\mathbf{s}^\perp|\mathbf{m}] = \frac{1}{2^{\mathcal{N}_v[\mathbf{m}]}}$$

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. We want here to determine $\mathcal{S}[\mathbf{m}]$ from (A1). By definition of the u_i^α we have

$$\mathbb{E}_f[m_\alpha] = 0 \qquad \text{and} \qquad \mathbb{E}_f[m_\alpha m_\beta] = \delta_{\alpha\beta},$$

so for large N_v we have

$$f[\mathbf{m}] = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{N_v}{2}\sum_{\alpha=1}^{d} m_\alpha^2\Big). \quad \text{(A2)}$$

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 [35]. With d assumed to be $\mathcal{O}(1)$, as $N_v \to \infty$ we expect in this regime a behaviour of the form

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

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 $f[\mathbf{m}]$. Denoting by $\boldsymbol{\mu} = \mathcal{O}(1)$ a conjugate d-dimensional vector and assuming that we can make sense to the following limit

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

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

$\mathcal{I}[\mathbf{m}]$ 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 $N_v \to \infty$)

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

$$= \mathbb{E}_{\boldsymbol{u}\sim p_{\boldsymbol{u}}}\Big[u^\alpha \tanh\Big(\sum_{\beta=1}^{d} u^\beta \mu_\beta\Big)\Big] \quad \text{(A3)}$$

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 (A2) 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: Effective Hamiltonian

We then expect to be able to rewrite the Hamiltonian corresponding to the visible distribution (3)

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

in terms of these new degrees of freedoms, with

$$\mathcal{H}[\mathbf{s}] = \sum_{i=1}^{N_v} \eta_i s_i - \sum_{j=1}^{N_h} \log\cosh\Big(\sum_{i=1}^{N_v} W_{ij}s_i - \theta_j\Big)$$

and

$$\mathcal{H}'[\mathbf{m}, \mathbf{s}^\perp] = N_v \sum_{\alpha=1}^{d} \eta_\alpha m_\alpha + N_v \sum_{\beta=d+1}^{N_v} \eta_\beta s_\beta[\mathbf{s}^\perp|\mathbf{m}]$$

$$- \sum_{j=1}^{N_h} \log\cosh\Big(\sqrt{N_v}\sum_{\alpha=1}^{d} w_\alpha m_\alpha v_j^\alpha$$

$$+ \sqrt{N_v}\sum_{\beta=d+1}^{N_v} w_\beta s_\beta[\mathbf{s}^\perp|\mathbf{m}]v_j^\beta - \theta_j\Big),$$

where $s_\beta[\mathbf{s}^\perp|\mathbf{m}]$ is a mapping to be defined of $\mathbf{s}^\perp$ to transverse projections $s_\beta = \frac{1}{\sqrt{N_v}}\sum_i s_i u_i^\beta$ given some magnetization $\mathbf{m}$, and $\eta_\alpha = \frac{1}{\sqrt{N_v}}\sum_i \eta_i u_i^\alpha$. We transform further

the RBM measure as

$$\begin{aligned} P[\mathbf{s}] &= P[\mathbf{m}, \mathbf{s}^\perp] \\ &= \frac{e^{-\mathcal{H}[\mathbf{s}^\perp|\mathbf{m}] - N_v \mathcal{F}_0[\mathbf{m}]}}{\sum_{\mathbf{s}^\perp} e^{-\mathcal{H}[\mathbf{s}^\perp|\mathbf{m}]}} \\ &= P(\mathbf{s}^\perp|\mathbf{m}) P(\mathbf{m}), \end{aligned}$$

with

$$P(\mathbf{m}) = \sum_{\mathbf{s}} P[\mathbf{s}] \prod_{\alpha=1}^{d} \delta(s_\alpha - m_\alpha) \stackrel{\text{def}}{=} e^{-N_v \mathcal{F}_0[\mathbf{m}]}.$$

We have introduced the Hamiltonian $\mathcal{H}\big[\mathbf{s}^\perp|\mathbf{m}\big]$ corresponding to the conditional measure $P(\mathbf{s}^\perp|\mathbf{m})$ while all the remaining dependency on $\mathbf{m}$ are stored (up to constant terms) in the quantity

$$\begin{aligned} \mathcal{F}_0[\mathbf{m}] = \sum_{\alpha=1}^{d} \eta_\alpha m_\alpha + \sum_{\beta=d+1}^{N_v} \eta_\beta m_\beta[\boldsymbol{\mu}] \\ - \frac{1}{N_v} \log\Big(\sum_{\mathbf{s}^\perp} e^{-\mathcal{H}[\mathbf{s}^\perp|\mathbf{m}]} \Big), \end{aligned}$$

where the transverse modes magnetizations $\beta > d$ are given by

$$m_\beta[\boldsymbol{\mu}] = \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} u_i^\beta \tanh\Big(\sqrt{N_v} \sum_{\alpha=1}^{d} \mu_\alpha u_i^\alpha \Big), \ \forall \beta > d \tag{B1}$$

as a function of $\{\mu_\alpha, \alpha = 1, \dots d\}$ solution to equation (A3) and are $\mathcal{O}(1/\sqrt{N_v})$. The transverse Hamiltonian corresponding to the probability of $\mathbf{s}^\perp$ conditional to $\mathbf{m}$ reads

$$\begin{aligned} \mathcal{H}\big[\mathbf{s}^\perp|\mathbf{m}\big] = \sum_{\beta=d+1}^{N_v} \eta_\beta (s_\beta[\mathbf{s}^\perp|\mathbf{m}] - m_\beta[\boldsymbol{\mu}]) \\ + \frac{1}{N_v} \sum_{j=1}^{N_h} \log\cosh\Big(\sqrt{N_v} \sum_{\alpha=1}^{d} w_\alpha m_\alpha v_j^\alpha \\ + \sum_{\beta=d+1}^{\min(N_v,N_h)} w_\beta s_\beta[\mathbf{s}^\perp|\mathbf{m}] v_j^\beta - \theta_j \Big), \end{aligned}$$

where $s_\beta[\mathbf{s}^\perp|\mathbf{m}]$ is a mapping of the configuration $\mathbf{s}^\perp$ onto $\mathbf{s}$ given some magnetization $\mathbf{m}$. The transverse configuration $(s_\beta, \beta = 1, \dots \min(N_v, N_h))$ satisfies

$$\sum_{\beta=d+1}^{N_v} s_\beta^2 = 1 - \sum_{\alpha=1}^{d} m_\alpha^2 \stackrel{\text{def}}{=} 1 - q^\parallel[\mathbf{m}]$$

and its entropy is $\mathcal{N}[\mathbf{m}] \log(2)$. At this point it is legitimate to introduce the set of spin variables $\mathbf{s}^\perp$ such that

$$\begin{aligned} s_{d+\gamma}[\mathbf{s}^\perp|\mathbf{m}] &= \sqrt{q^\perp[\mathbf{m}]} u_\gamma^{\perp 0} + r[\mathbf{m}] \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} s_\ell^\perp u_\gamma^{\perp \ell}, \\ &= m_{d+\gamma}[\boldsymbol{\mu}] + r[\mathbf{m}] \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} s_\ell^\perp u_\gamma^{\perp \ell}, \end{aligned}$$

for each $\gamma = 1, \dots \min(N_v, N_h) - d$, where the first term is a bias due to transverse magnetization $m_\beta, \beta > d$ while the second term represents the residual fluctuations. $\{u^{\perp \ell}, \ell = 0, \dots \mathcal{N}[\mathbf{m}]\}$ is an orthogonal set of $\mathcal{N}[\mathbf{m}] + 1$, normalized vectors decomposed onto the set $\{\boldsymbol{u}^\beta, \beta = d+1, \dots \min(N_v, N_h)\}$ as

$$\begin{aligned} \boldsymbol{u}^{\perp \mathbf{0}} &= \frac{1}{\sqrt{q^\perp[\mathbf{m}]}} \sum_{\beta=d+1}^{N_v} m_\beta \boldsymbol{u}^\beta \\ \boldsymbol{u}^{\perp \boldsymbol{\ell}} &= \sum_{\beta=1}^{\min(N_v,N_h)-d} u_\beta^{\perp \ell} \boldsymbol{u}^{\beta+d}, \ \ell = 1, \dots \mathcal{N}[\mathbf{m}], \end{aligned}$$

and the $s_\ell^\perp = 2B(1/2) - 1$ are iid spin variables (prior distribution), while we have introduced

$$q^\perp[\mathbf{m}] = \sum_{\beta=d+1}^{\min(N_v,N_h)} m_\beta^2 \tag{B2}$$

$$r[\mathbf{m}] = \sqrt{\frac{1 - q[\mathbf{m}]}{\mathcal{N}[\mathbf{m}]}}. \tag{B3}$$

With $\eta_\ell^\perp \stackrel{\text{def}}{=} \sqrt{N} \sum_{\beta=K+1}^{N} \eta_\beta u_{\beta-K}^\perp = \mathcal{O}(1)$, we get

$$\begin{aligned} \mathcal{H}\big[\mathbf{s}^\perp|\mathbf{m}\big] = \sqrt{N_v} r[\mathbf{m}] \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} \eta_\ell^\perp s_\ell^\perp \\ + \frac{1}{N_v} \sum_{j=1}^{N_h} \log\cosh\Big(\sqrt{N_v} \sum_{\alpha=1}^{\min(N_v,N_h)} w_\alpha m_\alpha v_j^\alpha \\ + r[\mathbf{m}] \sum_{\beta=d+1,\ell=1}^{\min(N_v,N_h),\mathcal{N}[\mathbf{m}]} w_\beta u_\beta^{\perp \ell} v_j^\beta s_\ell^\perp - \theta_j \Big). \end{aligned}$$

Finally, expanding up to second order the $\log\cosh$ we may obtain an effective disordered Ising model for the $\mathbf{s}^\perp$ variables:

$$\mathcal{H}_{\text{eff}}\big[\mathbf{s}^\perp|\mathbf{m}\big] = \sum_{\ell=1}^{\mathcal{N}[\mathbf{m}]} \eta_\ell^\perp[\mathbf{m}] s_\ell^\perp + \sum_{\ell,\ell'=1}^{\mathcal{N}[\mathbf{m}]} W_{\ell\ell'}^\perp[\mathbf{m}] s_\ell^\perp s_{\ell'}^\perp.$$

Introducing the notations:

$$\bar{m}_j = \tanh\Big(\sqrt{N_v}\sum_{\alpha=1}^{N_v} w_\alpha m_\alpha v_j^\alpha - \theta_j\Big),$$

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

we obtain

$$\eta_\ell^\perp[\mathbf{m}] = \sqrt{N_v} r[\mathbf{m}]\Big[\eta_\ell^\perp + \sum_{\beta=d+1}^{N_v}\bar{m}_\beta w_\beta u_\beta^{\perp\ell}\Big]$$

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

where $\eta_\ell^\perp[\mathbf{m}]$ is potentially $\mathcal{O}(1)$ while $W_{\ell,\ell'}^\perp[\mathbf{m}]$ is $\mathcal{O}\Big(\frac{1}{\sqrt{N_v}}\Big)$. The zero order term in the expansion of $\mathcal{H}\big[\mathbf{s}^\perp|\mathbf{m}\big]$ provides an additional contribution (7)

$$\mathcal{F}[\mathbf{m}] = \mathcal{F}_0[\mathbf{m}] - V[\mathbf{m}]$$

to the free energy which we further decompose into transverse and longitudinal free energy $\mathcal{F}^\parallel$ and $\mathcal{F}^\perp$ given in (5,6). To make connection with data, i.e. given a configuration $\mathbf{s}$ with magnetization $m_\alpha, \alpha = 1,\dots K$, the $\mathbf{s}^\perp$ are constructed as follows. First let

$$m_\ell^\perp[\mathbf{s}] = \frac{1}{1-q^\parallel[\mathbf{m}]}\sum_{\beta=K+1}^{N} s_\beta u_\beta^{\perp\ell} \in [-1,1],$$

for each $\ell = 1,\dots,\mathcal{N}[\mathbf{m}]$ the magnetization of the configuration $\mathbf{s}$ along this mode. This allows us to define the probability

$$p_\ell[\mathbf{s}] = \frac{1+m_\ell^\perp[\mathbf{s}]}{2} \in [0,1].$$

Then with

$$s_\ell^\perp = 2B\big(p_\ell[\mathbf{s}]\big) - 1, \qquad \forall \ell = 1,\dots\mathcal{N}[\mathbf{m}]$$

with $B(p)$ the Bernoulli distribution of parameter p, we have a set of spin variables fulfilling our needs.

Appendix C: Coulomb interaction picture

We can rewrite $V[\mathbf{m}]$ as

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

where the inverse d-dimensional Laplace kernel $K_d(|m-m'|)$ is by definition

$$\Delta_d K_d(|\mathbf{m}-\mathbf{m}'|) = \delta(\mathbf{m}-\mathbf{m}').$$

As a result, by construction $V[\mathbf{m}]$ is solution to the Poisson equation

$$\Delta_d V[\mathbf{m}] = \rho(\mathbf{m})$$

with the density of Coulomb charges given by

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

To make sense of this quantity we 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

$$\rho(\mathbf{m}) = \frac{2}{N_v}\sum_{j=1}^{N_h}\nu_j\delta_{\nu_j}\big(\mathbf{n}_j^T\mathbf{m} - \tilde{\theta}_j\big)$$

with ν_j and n_j given by equation (10,11) and $\tilde{\theta}_j = \theta_j/\nu_j$. In this form, ρ is readily a superposition of N_h uniformly charged hyperplanes. Each hyperplane j being defined by a normal vector $\mathbf{n}_j$, an offset $\tilde{\theta}_j$ from the origin, a width ν_j^{-1} and a (hyper)surface charge density $2\nu_j/\sqrt{N_v}$. Furthermore, integrating the d-dimensional Poisson kernel over transverse variable $\mathbf{m}'^\perp$ w.r.t. some unit vector $\mathbf{n}$ at distance θ of the origin yields the 1-dimensional one:

$$\int d\mathbf{m}'^\perp K_d\big(|\mathbf{m}-\mathbf{m}'|\big) = |\mathbf{n}^T\mathbf{m} - \theta|.$$

As a result the one particle potential takes the form

$$V[\mathbf{m}] = \frac{2}{N_h}\sum_{j=1}^{N_h}\nu_j\int d\theta\rho_{\nu_j}(\theta)|\mathbf{n}_j^T\mathbf{m}-\theta|.$$

Appendix D: 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

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

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

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

Indeed, let

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

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 $\frac{\nu}{2}\big[1-\tanh(\nu x)\big]$ 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

$$V(m) = -\frac{m^2}{2N_v\epsilon} + \eta m + \sum_k \rho_k|m-\theta_k|$$

with

$$\rho_k = \frac{1}{2N_v\epsilon}(m_{k+1}-m_k),$$

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 E: RBM optimization seen as a linear regression

The projection of the empirical distribution onto the space of RBM 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

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

with

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

where $\hat{\rho}(\mathbf{m})$ is 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 $\Theta(\mathbf{n},\theta)$ yielding a free energy of the form

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

with

$$\rho(\mathbf{m}) = \sum_{j=1}^{N_h}\Theta_j\delta\big(\mathbf{n}_j^T\mathbf{m}-\theta_j\big).$$

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

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

between $\hat{\rho}$ and ρ is minimized, J being the Fisher metric defined as

$$J\big[\mathbf{m}_1, \mathbf{m}_2\big] = \mathrm{Cov}_{\mathbf{m}\sim p(\mathbf{m}|\Theta)}\Big[K_d(|\mathbf{m}-\mathbf{m}_1|), K_d(|\mathbf{m}-\mathbf{m}_2|)\Big],$$
$$\simeq \mathrm{Cov}_{\mathbf{m}\sim \hat{p}(\mathbf{m})}\Big[K_d(|\mathbf{m}-\mathbf{m}_1|), K_d(|\mathbf{m}-\mathbf{m}_2|)\Big].$$

As we shall see this projection turns out to be a linear regression of the centered random variable

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

onto the set of centered random variables (the score variables associated to Θ)

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

$\mathbb{E}_{\mathbf{m}\sim\hat{p}(\mathbf{m})}$ 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

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

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 projection is given by $V^{\parallel}$ 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 un-tractable since $\hat{V}$ involves a very complicated density of charge $\hat{\rho}$. This is not the case because by construction we have

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

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

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

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

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

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 Figure 4. In this appealing picture there is however a loophole hidden under the carpet. 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 would seem to provide some 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. This requires the linear regression to be complemented with some additional regularization which remains out of our reach at this point.

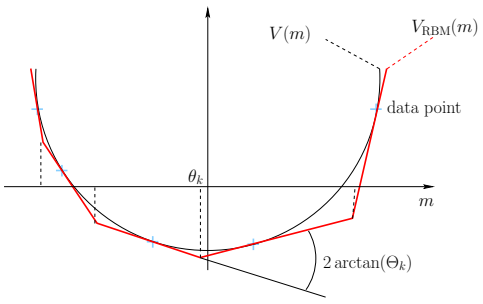

FIG. 4. Picture of the $V_{\mathrm{RBM}}(\mathbf{m})$ (in blue) at the interpolation threshold in 1-d.