IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **53** (2020) 094001 (18pp) https://doi.org/10.1088/1751-8121/ab6b92

Legendre equivalences of spherical Boltzmann machines

Giuseppe Genovese[1] **and Daniele Tantari**[2]

[1] Department of Mathematics and Computer Science, University of Basel, Spiegelgasse 1, 4051 Basel, Switzerland
[2] Mathematics Department, University of Bologna, Via Zamboni, 33, 40126, Bologna, Italy

E-mail: giuseppe.genovese@unibas.ch and daniele.tantari@unibo.it

Received 31 October 2019, revised 30 December 2019
Accepted for publication 14 January 2020
Published 4 February 2020

Abstract

We study either fully visible and restricted Boltzmann machines with sub-Gaussian random weights and spherical or Gaussian priors. We prove that the free energies of the spherical and Gaussian models are related by a Legendre transformation. Incidentally our analysis brings also a new purely variational derivation of the free energy of the spherical models.

Keywords: Boltzmann machines, Hopfield model, Legendre equivalence, free energy, soft spins

1. Introduction

Originally inspired by statistical physics [1], Boltzmann machines (BMs) [2, 3] are among the most studied data generative models, playing a central role in the phenomenal progresses of machine learning through neural networks of the last two decades. In particular restricted BMs (RBMs) constitute a cornerstone of unsupervised learning, mainly for the very successful training algorithms developed [4, 5], working also for many interesting deep architectures [6, 7], for which RBMs are used as the basic building blocks [8, 9].

Concretely a BM is a probability distribution of the Gibbs type which is aimed to reproduce the true distribution of the data. In the much useful neural network interpretation the units of the machine should mirror the data, that is typical configurations according to the BM distribution are desired to be close to typical data. Therefore two ingredients are crucial to build up a BM: the energy function and the *a priori* unit distribution. The main focus of the paper will be on fully visible BMs, namely Hopfiel models, and RBMs with spherically symmetric priors. We will give a formula for the free energy of these BMs pointing out a Legendre duality between rigid spherical priors and a certain quite general class of sub-Gaussian distributions, already investigated for the Sherrington–Kirkpatrick energy [10]. This latter equivalence is

achieved by a suitable adaption of a very much established method from the statistical physics tradition, namely equivalence of ensemble (spherical and Gaussian).

1.1. Set up

First we will introduce the models we will deal with. We start by the energy function.

Let $\{\xi_{ij}\}_{i=1,\dots,N_1,j=1,\dots,N_2}$ a doubly indexed sequence of i.i.d. centred sub-Gaussian r.vs. with

$$E[\xi_{ij}\xi_{hk}] = \delta_{ih}\delta_{jk}.$$

For definiteness we assume that

$$\frac{N_1}{N_1+N_2} \to \alpha \in (0,1).$$

We shall look at this sequence in two different ways, namely as entries of a $N_1 \times N_2$ random matrix Ξ or a collection of N_2 patterns in $\mathbb{R}^{N_1}$, defining a sample covariance matrix $\frac{1}{N_2}\Xi\Xi^T \in \mathbb{R}^{N_1\times N_1}$. In either cases we can use the following two important properties of rectangular random matrices with centred independent subgaussian entries with unitary variance [11, 12]. For $A \in \mathbb{R}^{N\times N}$ we denote its eigenvalues as $\lambda_i := \lambda_i(A)$, $i = 1,\dots N$; for $A \in \mathbb{R}^{N_1\times N_2}$ we denote its singular values as $\sigma_i := \sigma_i(A)$, $i = 1,\dots N_1$.

(**P1**) The empirical distribution of the eigenvalues of $\frac{1}{N_2}\Xi\Xi^T$ converges a.s. to the Marchenko–Pastur law:

$$\frac{1}{N_1}\sum_{j=1}^{N_1}\delta(\lambda-\lambda_i(\frac{1}{N_2}\Xi\Xi^T)) \rightharpoonup \rho_{\mathrm{MP}}(\lambda;\alpha) \quad \mathbb{P}-\text{a.s.}$$

where

$$\rho_{\mathrm{MP}}(\lambda;\alpha) := (2-\frac{1}{\alpha})^+\delta_0(\lambda) + \frac{(1-\alpha)}{2\pi\alpha}\frac{\sqrt{(\lambda-\lambda_-)(\lambda_+-\lambda)}}{\lambda}1_{[\lambda_-,\lambda_+]}(\lambda) \tag{1.1}$$

with $\lambda_\pm := \left(1\pm\sqrt{\alpha/(1-\alpha)}\right)^2$.

(**P2**) The spectrum of $\frac{1}{N_2}\Xi\Xi^T$ is localised in an interval with large probability:

$$\mathbb{P}(\|\frac{1}{N_2}\Xi\Xi^T - \mathbb{I}\|_{op} \geqslant t+\sqrt{\lambda_+}-1) \leqslant 2\mathrm{e}^{-\frac{N_1t^2}{2}}. \tag{1.2}$$

We will deal with two kind of Boltzmann machines: Hopfield models (HMs) and restricted Boltzmann machines (RBMs). Their energy functions (or Hamiltonians) are

$$H^{\mathrm{HM}}_{N_1,N_2} := -\frac{1}{N_1+N_2}\sum_{j=1}^{N_2}\sum_{i<k}^{N_1}\xi_{ij}\xi_{kj}x_ix_k, \tag{1.3}$$

$$H^{\mathrm{RBM}}_{N_1,N_2} := -\frac{1}{\sqrt{N_1+N_2}}\sum_{j=1}^{N_2}\sum_{i=1}^{N_1}\xi_{ij}x_iy_j \tag{1.4}$$

(we will often drop all the indexes from the energies to lighten the notations).

In the Hopfield model units have one single choice for the prior distribution, while a RBM is an undirected bipartite system in which we can have different priors for each layer. With this in mind, we can now introduce the prior distributions we shall deal with. Let $S^N(R)$ be the $(N-1)$-dimensional sphere in $\mathbb{R}^N$ with radius $R\sqrt{N}$. Define the following *a priori* probability measures on $\mathbb{R}^N$:

$$\begin{cases} \sigma_{N,R}(\mathrm{d}x) & \text{uniform on } S^N(R); \\ \gamma_{N,\theta}(\mathrm{d}x) & \text{centred Gaussian with covariance } \theta\mathbb{I}. \end{cases} \tag{1.5}$$

Models with Gaussian priors are typically ill-defined for low temperatures and need a sub-Gaussian regularisation. Let $r : \mathbb{R} \mapsto \mathbb{R}$ such that there is $\varepsilon > 0$ with

$$\lim_{x\to\infty} \frac{r(x)}{x^{2+\varepsilon}} = \infty.$$

We define the regularised prior on $\mathbb{R}^N$

$$\rho_N(\mathrm{d}x) := \mathrm{e}^{-Nr\left(\frac{\|x\|}{\sqrt{N}}\right)}\gamma_{N,\theta}(\mathrm{d}x). \tag{1.6}$$

For instance in [13] it was considered $r(x) = \beta x^4/4$ while in [10, 14] $r(x) = \beta x^4/4 - \lambda x^2/2$.

This notion extends easily to two layer settings. We introduce some $r : \mathbb{R}^2 \mapsto \mathbb{R}$ so that

$$\lim_{x^2+y^2\to\infty} \frac{r(x,y)}{x^2} = \lim_{x^2+y^2\to\infty} \frac{r(x,y)}{y^2} = \infty.$$

Then we define the following measure on $\mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$

$$\rho^2_{N_1,N_2}(\mathrm{d}x\mathrm{d}y) := \mathrm{e}^{-\sqrt{N_1N_2}r\left(\frac{\|x\|}{\sqrt{N_1}},\frac{\|y\|}{\sqrt{N_2}}\right)}\gamma_{N_1,\theta}(\mathrm{d}x)\gamma_{N_2,\theta}(\mathrm{d}y). \tag{1.7}$$

The idea is that ρ can be used to regularise a single layer, while ρ^2 regularises two layers at once. A simple example is low rank matrix factorisation with Gaussian priors in which $r(x,y) = x^2y^2/2$ [15].

One technical problem is that the support of $\sigma_{N,R}$ has zero $\gamma_{N,\theta}$-measure, that is $\gamma_{N,\theta}(S^N(R)) = 0$. For this reason for a given $\varepsilon > 0$ we need to introduce the spherical shells

$$S_{N,R,\varepsilon} := \{x \in \mathbb{R}^N : R - \varepsilon \leqslant \|x\|_2 \leqslant R + \varepsilon\}. \tag{1.8}$$

We denote (here and further $|A|$ is the Lebesgue measure of the set A)

$$\sigma_{N,R,\varepsilon}(\mathrm{d}x) := |S_{\varepsilon,N,R}|^{-1}1_{\{S_{\varepsilon,N,R}\}}(x)\mathrm{d}x$$

the uniform probability on a shell and note that it is a.c. w.r.t. $\gamma_{N,\theta}$. We will also often make use of the simple relation

$$|S_{\varepsilon,N,R}| = \varepsilon|S_{N,R}| + O\left(\varepsilon^2\right). \tag{1.9}$$

With these definitions at hand, we can introduce the probability distributions defining our Boltzmann machines. Let $\mu_{(\cdot)} \in \{\sigma_{(\cdot)}, \gamma_{(\cdot)}, \rho_{(\cdot)}\}$ and $\mu^2_{(\cdot)} \in \{\sigma_{(\cdot)}, \gamma_{(\cdot)}\} \otimes \{\sigma_{(\cdot)}, \gamma_{(\cdot)}\} \cup \{\rho^2_{(\cdot)}\}$ denote one prior among the one introduced before respectively for the one-layer and the bipartite machine. We have for $\beta > 0$

$$G^{\mathrm{HM}}_{N_1,N_2,\beta}(\mathrm{d}x) := Z^{-1}_{N_1,N_2,\beta}\mu_{N_1}(\mathrm{d}x)\mathrm{e}^{-\beta H(x)} \qquad G^{\mathrm{RBM}}_{N_1,N_2,\beta} := Z^{-1}_{N_1,N_2,\beta}\mu^2_{N_1,N_2}(\mathrm{d}x\mathrm{d}y)\mathrm{e}^{-\beta H(x,y)}. \tag{1.10}$$

The normalisation $Z_{N_1,N_2,\beta}$ is called partition function and it needs not to be the same, despite the symbol. Moreover

$$A^{\mathrm{HM}}_{N_1,N_2}(\beta) := \frac{1}{N_1+N_2}\log Z_{N_1,N_2,\beta} \qquad A^{\mathrm{RBM}}_{N_1,N_2}(\beta) := \frac{1}{N_1+N_2}\log Z_{N_1,N_2,\beta}. \tag{1.11}$$

Most interesting is to evaluate the last quantities in the so-called thermodynamic limit, namely

$$N_1, N_2 \to \infty, \quad \text{such that} \qquad \lim_{N_1,N_2} \frac{N_1}{N_1+N_2} =: \alpha > 0.$$

The regime $\alpha \neq 0, 1$ is called high-load and we will stick to that in this work.

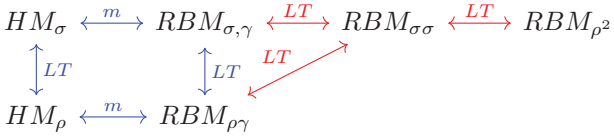
Being Lipschitz functions of the weights, free energies always satisfies a concentration inequality which ensures their a.s. convergence to the expected value. This self-averaging property will be exploited throughout without further mention.

For simplicity we will assume always the distributions of the two layers to have the same parameters, i.e. radius and variance. The general case requires a trivial extension of our formulas.

In general $A_\mu(X, \alpha, \beta)$ denotes the free energy of a BM with prior μ whose parameters are X. We will often drop the descriptive labels from the Hamiltonian, Gibbs measure and free energy when it will be clear from the context to which one we refer to. Only exception, the quantities of interest referred to spherical shell priors are indicated by $\hat{\cdot}$ for all the models.

1.2. Main result and organisation of the paper

We will focus on the free energy associated to BMs with the particular priors introduced above. We will prove the following equivalence at the level of free energies, which can be related by a marginalisation (m) or a Legendre transform (LT):



We will not concern here about low-load ($\alpha = 0, 1$), yet some of the equivalences we state hold also in this regime. More precisely, red arrows indicate equivalences valid only in high load while blue arrow equivalences hold regardless of α.

Marginalisation is the usual trick of RBMs. The two layers are coupled linearly, so that one can integrate out á la Stratonovich the units from the Gaussian layer in the partition function of a RBM to obtain the partition function of a HM (with β^2 replacing β). All the relevant quantities (e.g. Gibbs measures, free energy, order parameters etc) of one model can be computed directly from the one of the other one. For instance the equivalence $\mathrm{HM}_\sigma \leftrightarrow \mathrm{RBM}_{\sigma,\gamma}$ at the level of Gibbs measure reads as

$$Z^{-1}_{N_1,N_2,\beta}\sigma_{N_2,R}(\mathrm{d}x)\int_{\mathbb{R}^{N_2}} \gamma_{N_2,\theta}(\mathrm{d}y)\mathrm{e}^{\beta H^{\mathrm{RBM}}_{N_1,N_2}} = Z^{-1}_{N_1,N_2,\beta}\sigma_{N_1,R}(\mathrm{d}x)\mathrm{e}^{\frac{\beta^2}{\theta^2}H^{\mathrm{HM}}_{N_1}},$$

since the partition functions of the two models are numerically the same.

Legendre transforms are where the idea of equivalence of ensembles exploits and more precise statements are given in theorem 1.1 below. A Gaussian prior of N units concentrates on a N-dimensional sphere of radius proportional to $\sqrt{N}$. To find the optimal radius we slice up the Gibbs measure at the level of the Gaussian prior and look for the most relevant contribution to the free energy. This strategy yields naturally a variational principle of the Legendre

type relating the spherical and Gaussian free energies. Moreover we identify the square radius of the optimal sphere R^2 and the variance of the Gaussian model θ^{-1} as Legendre conjugate variables.

The main idea is very simple; we briefly outline the heuristics for the equivalence $\mathrm{RBM}_{\sigma\sigma} \overset{\mathrm{LT}}{\longleftrightarrow} \mathrm{RBM}_{\rho^2}$ (the other cases are similar). By the standard disintegration of finite-dimensional Gaussian measures into spheres we have

$$\begin{aligned}
&\int_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_1}} \mu^2(\mathrm{d}x\mathrm{d}y)\mathrm{e}^{-\beta H(x,y)} \\
&= \int_0^\infty \mathrm{d}R_1 \int_0^\infty \mathrm{d}R_2 \mathrm{e}^{-\sqrt{N_1N_2}r\left(\frac{R_1}{\sqrt{N_1}},\frac{R_2}{\sqrt{N_2}}\right)-\frac{R_1}{2\theta_1}-\frac{R_2}{2\theta_2}} \\
&\frac{|S_{N_1}(R_1)\times S_{N_2}(R_2)|}{\sqrt{2\pi\theta}^{N_1+N_2}} \int_{S_{N_1}(R_1)\times S_{N_2}(R_2)} \sigma_{N_1,R_1}(\mathrm{d}x)\sigma_{N_2,R_2}(\mathrm{d}y)\mathrm{e}^{-\beta H(x,y)}.
\end{aligned} \tag{1.12}$$

We adjusted the normalisation of the inner integral so to get the partition function of $\mathrm{RBM}_{\sigma\sigma}$. Thus we continue the chain of identities as

$$(1.12) = \int_0^\infty \mathrm{d}R_1 \int_0^\infty \mathrm{d}R_2 \mathrm{e}^{-\sqrt{N_1N_2}r\left(\frac{R_1}{\sqrt{N_1}},\frac{R_2}{\sqrt{N_2}}\right)-\frac{R_1}{2\theta_1}-\frac{R_2}{2\theta_2}+\log\left(\frac{|S_{N_1}(R_1)\times S_{N_2}(R_2)|}{\sqrt{2\pi\theta}^{N_1+N_2}}\right)+(N_1+N_2)A_{\sigma\sigma}(R_1,R_2)}. \tag{1.13}$$

Then we can evaluate the integral by the usual Laplace method, since by a simple scaling argument the maximum must be attained at the scale $R_1, R_2 \sim \sqrt{N_1+N_2}$.

To give rigorous grounds to this heuristics we need some few properties. First, we have to control the thermodynamic limit of the free energy of spherical models HM_σ and $\mathrm{RBM}_{\sigma\sigma}$. These limits are computed in section 2. Secondly, to properly implement the above disintegration formula, we need the regular behaviour of the model with spherical shell prior as the thickness of the shell vanish. In other words, at the level of free energy thermodynamics should favour those configurations on the shell which actually lie on a given sphere into it. This is proven in sections 2 and 3. Lastly, the unit configurations outside any ball of radius growing faster than $\sqrt{N_1}, \sqrt{N_2}$ must give vanishing contribution in the thermodynamic limit. This is proven in section 3, where the proof of our main result, i.e. subsequent theorem 1.1, is completed.

Theorem 1.1. *Assume* (**P1**), (**P2**) *and let* ρ_N, ρ_N^2 *be defined as in (1.6), (1.7) and discussion around. Then*

(i) $\mathrm{HM}_\sigma \overset{\mathrm{LT}}{\longleftrightarrow} \mathrm{HM}_\rho$*:*

$$A_\rho(\theta,\alpha,\beta) = \alpha \sup_{R>0}\left(\alpha^{-1}A_\sigma(R,\beta) - \frac{R^2}{2\theta} + \log R - r(R) - \frac{1}{2}(\log\theta - 1)\right); \tag{1.14}$$

$$A_\sigma(R,\beta) = \alpha \inf_{\theta>0}\left(\frac{R^2}{2\theta} + \alpha^{-1}A_\rho(\theta,\alpha,\beta) - \log R + r(R) + \frac{1}{2}(\log\theta - 1)\right). \tag{1.15}$$

(ii) $RBM_{\sigma\sigma} \overset{\mathrm{LT}}{\longleftrightarrow} \mathrm{RBM}_{\sigma,\gamma}$*:*

$$A_{\sigma,\gamma}(\theta,R_2,\alpha,\beta) = \alpha \sup_{R_1>0}\left(\alpha^{-1}A_{\sigma\sigma}(R_1,R_2,\beta) - \frac{R_1^2}{2\theta} + \log R_1 - \frac{1}{2}(\log\theta - 1)\right); \tag{1.16}$$

$$A_{\sigma\sigma}(R_1,R_2,\beta)=\alpha\inf_{\theta>0}\left(\frac{R_1^2}{2\theta}+\alpha^{-1}A_{\sigma,\gamma}(\theta,R_2,\alpha,\beta)-\log R_1+\frac{1}{2}(\log\theta-1)\right). \tag{1.17}$$

(iii) RBM$_{\sigma\sigma}$ $\overset{\text{LT}}{\longleftrightarrow}$ *RBM$_{\rho^2}$:*

$$A_{\rho^2}(\theta,\alpha,\beta)=\sup_{R_1,R_2>0}\left(A_{\sigma\sigma}(R_1,R_2)-\frac{\alpha R_1^2+(1-\alpha)R_2^2}{2\theta}+\log(R_1^{\alpha}R_2^{1-\alpha})-\sqrt{\alpha(1-\alpha)}r(R_1,R_2)-\frac{1}{2}(\log\theta-1)\right); \tag{1.18}$$

$$A(R_1,R_2)=\inf_{\theta_1,\theta_2>0}\left(\frac{\alpha R_1^2}{2\theta_1}+\frac{(1-\alpha)R_2^2}{2\theta_2}+A_{\rho^2}(\theta_1,\theta_2,\alpha,\beta)-\log(R_1^{\alpha}R_2^{1-\alpha})+\sqrt{\alpha(1-\alpha)}r(R_1,R_2)+\frac{1}{2}(\log\theta_1^{\alpha}\theta_2^{1-\alpha}-1)\right). \tag{1.19}$$

The other equivalences of the scheme above can be derived by combining (i)–(iii) and marginalisations. Albeit we will not include that in this paper, these dualities can be established also for the Gibbs distributions (1.10).

1.3. Related literature

Once the model has been properly regularised, spherical, Gaussian or sub-Gaussian priors are all equivalent; so we shall speak indistinctly of Gaussian BMs in what follows.

The use of Gaussian visible variables is useful to handle real data and has been suggested since the beginning of the theory [16]. However the learning and retrieval capabilities of the fully visible BM with Gaussian units are not as good as its ±1 counterpart at low-load [17], and at high-load they are totally useless [17, 18]. Restricted architectures are more interesting. RBMs with Gaussian visible and latent variables have been used for instance for factor analysis [19, 20] and collaborative filtering [21]. In general training, through e.g. contrastive divergence, is slower than that in a Bernoulli–Gaussian machine [22–25] and also retrieval is less pronounced [17, 26].

Independently on their performances, Gaussian BMs are of great theoretical relevance from the viewpoint of spin glasses, since their simpler mathematical structure helps our understanding of the, much more complicated, discrete models. Previous results on the model have been obtained in [27] and [28]. In [28] the authors achieve the same result as in our theorem 2.2 and prove much more: fluctuations of the free energy are shown to be Gaussian in high temperature and Tracy–Widom for low temperature. The assumptions on the weight distribution are more general then ours, as they only require finiteness of the moments. Yet the method there employed is a sophisticated and technical random matrix argument and our approach is certainly lighter and more accessible to non-specialists. In [27] a variational principle for the free energy has been proven only for small β, by means of the so-called Latala method. One great merit of the approach of this paper is to provide a clear interpretation of the replica symmetric nature of the variational formula for the free energy (formulated in terms of the overlap), which is absent in [28] and in the present work, even though by a direct comparison with [10] one can see that our Lagrange multipliers $(a, b$ below) are essentially shifted overlaps. In any case it is remarkable that the free energy of a doubly spherical RBM satisfies a fully convex minimisation principle, which is a crucial difference compared to the min max of Gauss–Bernoulli [29] and Bernoulli–Bernoulli RBMs [30]. The reason for that eludes our current understanding and we must defer the discussion of this point to future works.

2. Free energy of spherical models

In this section we study the HM Hamiltonian (1.3) and RBM Hamiltonian (1.4) with the spherical prior in (1.5). Our main results are

Theorem 2.1. *Let A_{N_1,N_2} denote the free energy of HM_σ. It holds P-a.s.*

$$\lim_{N_1,N_2\to\infty} A_{N_1,N_2} = \alpha \min_{2q\geqslant\beta(1-\alpha)\lambda_+} \left(qR^2 - \frac{1}{2}\int \rho_{\mathrm{MP}}(\lambda;\alpha)\log(2q-\beta(1-\alpha)\lambda)\mathrm{d}\lambda - \log R - \frac{1}{2}\right). \tag{2.1}$$

Theorem 2.2. *Let A_{N_1,N_2} denote the free energy of $RBM_{\sigma\sigma}$ and $\alpha \leqslant \frac{1}{2}$. It holds P-a.s.*

$$\lim_{N_1,N_2\to\infty} A_{N_1,N_2} = \min_{ab\geqslant\beta^2(1-\alpha)\lambda_+} \left(\frac{R^2}{2}(\alpha a + (1-\alpha)b) - \frac{1}{2}(1-2\alpha)\log(b) \right.$$
$$\left. - \frac{\alpha}{2}\int \rho_{\mathrm{MP}}(\lambda;\alpha)\log(ab-\beta^2(1-\alpha)\lambda)\mathrm{d}\lambda - \log R - \frac{1}{2}\right). \tag{2.2}$$

The choice $\alpha \leqslant \frac{1}{2}$ is just a matter of convenience as it will be clear that pre-choosing the largest layer simplifies a lot the notations. For $\alpha \geqslant \frac{1}{2}$ one should bear in mind that the Marchenko–Pastur distribution (1.1) has an atom in zero.

First of all we prove that the spherical shells constitute a good approximation of the spherical prior. We recall that everywhere the quantities with $\widehat{\cdot}$ are always referred to spherical shell priors.

Lemma 2.3. *Let $\widehat{A}_{N_1,N_2,\varepsilon}$ denote the free energy of HM_{σ_ε}, $RBM_{\sigma_\varepsilon\sigma}$ or $RBM_{\sigma_\varepsilon\sigma_\varepsilon}$. Then it is*

$$\lim_{\varepsilon\to 0}\lim_{N_1,N_2} \widehat{A}_{N_1,N_2,\varepsilon} = \lim_{N_1,N_2}\lim_{\varepsilon\to 0} \widehat{A}_{N_1,N_2,\varepsilon}.$$

The existence of the limit in the r.h.s. will be proven below.

Proof. In this proof we write N to mean N_1 or N_2 and by $A^\sigma(R;\beta)$ the free energy of a spherical model of radius R, according to the context. No further details are needed. By Fubini and the mean value theorem there is a $R^*_\varepsilon \in [R\sqrt{N}-\varepsilon, R\sqrt{N}+\varepsilon]$ for which

$$\widehat{A}_{N_1,N_2} = A^\sigma_{N_1,N_2}(R^*_\varepsilon;\beta).$$

By a simple change of variables we have

$$A^\sigma_{N_1,N_2}(R^*_\varepsilon;\beta) = A^\sigma_{N_1,N_2}(R;c_{N,\varepsilon}\beta),$$

with $c_{N,\varepsilon} := R^2_\varepsilon/R^2N$. Therefore by Lipschitz continuity of the free energy w.r.t. β

$$|A^\sigma_{N_1,N_2} - \widehat{A}_{N_1,N_2,\varepsilon}| \leqslant \frac{\varepsilon}{N}.$$

This, combined with $A^\sigma_{N_1,N_2} = \lim_{\varepsilon\to 0}\widehat{A}_{N_1,N_2,\varepsilon}$, gives the assert. □

We first deal with the Hopfield model. Everywhere from now on partition function and pressure will be referred to this model, unless otherwise specified.

The best advantage of the spherical prior is that one can diagonalise the energy (1.3):

$$H(x) = -(1-\alpha)\sum_{i=1}^{N_1}\lambda_i x_i^2.$$

Thanks to **(P2)**, we can restrict our analysis to disorder realisations with spectrum contained in $(-\infty, \lambda_+]$. More precisely

Lemma 2.4. *For any $a > \lambda_+$ there is a $c > 0$ such that*

$$E[A_{N_1,N_2}(\beta,R)1_{\{\|\frac{\Xi\Xi^T}{N_2}\|_{op}>a\}}] = O\left(\mathrm{e}^{-cN_1}\right). \tag{2.3}$$

The proof of that follows essentially the same lines of [10]. We omit here the details. Also, the next proof is a straightforward adaption of [10, proof of (9)]. We give it completely in order to introduce the argument used to prove theorem 2.2.

Proof of theorem 2.1. Let $2q > \beta(1-\alpha)\lambda_+$. Then using (1.9) we have

$$\begin{aligned}\varepsilon\hat{Z}_{N_1,N_2,\varepsilon}(\beta,\alpha,R) &= \varepsilon\,\mathrm{e}^{qR^2N_1}\int_{\mathbb{R}^{N_1}}\sigma_{N_1,R,\varepsilon}(\mathrm{d}z)\mathrm{e}^{-\sum_i^{N_1}(q-\frac{\beta(1-\alpha)}{2}\lambda_i)z_i^2}\\ &\leqslant \mathrm{e}^{qR^2N_1}\frac{(2\pi)^{\frac{N_1}{2}}}{|S_{R\sqrt{N_1}}|}\int_{\mathbb{R}^N}\frac{\mathrm{d}z}{(2\pi)^{\frac{N_1}{2}}}\mathrm{e}^{-\sum_i^{N_1}(2q-\beta(1-\alpha)\lambda_i)z_i^2/2}\\ &= \mathrm{e}^{qR^2N_1}\frac{(2\pi)^{\frac{N_1}{2}}}{|S_{R\sqrt{N}}|}\mathrm{e}^{-\frac{1}{2}\sum_i^N\log(2q-\beta(1-\alpha)\lambda_i)},\end{aligned} \tag{2.4}$$

therefore, since $\frac{1}{N_1}\log\left(|S_{R\sqrt{N_1}}|/\sqrt{2\pi}^{N_1}\right) \to \log R + \frac{1}{2}$ and thanks to lemma 2.3

$$\limsup_{N_1,N_2}\frac{1}{N_1+N_2}\log Z_{N_1,N_2}(\beta,\alpha,R) \leqslant \alpha q R^2 - \frac{\alpha}{2}\int\rho_{\mathrm{MP}}(\lambda;\alpha)\log(2q-\beta(1-\alpha)\lambda) - \alpha\log R - \frac{\alpha}{2} =: \tilde{A}(q),$$

whence, as $\tilde{A}(q)$ is continuous,

$$\limsup_{N_1,N_2}\frac{1}{N_1+N_2}\log Z_{N_1,N_2}(\beta,\alpha,R) \leqslant \min_{2q\geqslant\beta(1-\alpha)\lambda_+}\tilde{A}(q).$$

Moreover for $2q > \beta(1-\alpha)\lambda_+$

$$\partial_q^2\tilde{A}(q) = 2\alpha\int\mathrm{d}\lambda\frac{\rho_{\mathrm{MP}}(\lambda;\alpha)}{(2q-\beta(1-\alpha)\lambda)^2} > 0,$$

thus $\tilde{A}(q)$ is convex and the minimum is attained in a unique point $\bar{q}$.

Now the reverse bound. Let $\varepsilon > 0$ and $S^c_{N_1,R,\varepsilon}$ the complementary set of the shell $S_{N_1,R,\varepsilon}$. It holds

$$\varepsilon\widehat{Z}_{N_1,N_2,\varepsilon} = \mathrm{e}^{qR^2N_1}\frac{(2\pi)^{\frac{N_1}{2}}}{|S_{R\sqrt{N_1}}|}\int_{\mathbb{R}^{N_1}}\frac{\mathrm{d}z}{(2\pi)^{\frac{N_1}{2}}}\mathrm{e}^{-\sum_i^{N_1}(q-\frac{\beta(1-\alpha)}{2}\lambda_i)z_i^2} - \mathrm{e}^{qR^2N_1}\frac{(2\pi)^{\frac{N}{2}}}{|S_{R\sqrt{N_1}}|}\int_{S^c_{N_1,R,\varepsilon}}\frac{\mathrm{d}z}{(2\pi)^{\frac{N_1}{2}}}\mathrm{e}^{-\sum_i^{N_1}(q-\frac{\beta(1-\alpha)}{2}\lambda_i)z_i^2}.$$

For any $\eta > 0$ small enough we have

$$\begin{aligned}\int_{S^c_{N_1,R,\varepsilon}}\frac{\mathrm{d}z}{(2\pi)^{\frac{N_1}{2}}}\mathrm{e}^{-\sum_i^{N_1}(q-\frac{\beta(1-\alpha)}{2}\lambda_i)z_i^2} &\leqslant \exp\left[N_1\left(\eta\left(R^2-\frac{\varepsilon}{N_1}\right)-\frac{1}{2N_1}\sum_j\log(2q-\beta\lambda_j+2\eta)\right)\right]\\ &+ \exp\left[N_1\left(-\eta\left(R^2+\frac{\varepsilon}{N_1}\right)-\frac{1}{2N_1}\sum_j\log(2q-\beta\lambda_j-2\eta)\right)\right].\end{aligned}$$

Note now that the r.h.s. is $o(\mathrm{e}^{-N})$ if $\frac{\varepsilon}{N} \to \infty$ as $N \to \infty$. So the greatest contribution is at the scale $\varepsilon = \tilde{\varepsilon}N$, $\tilde{\varepsilon} > 0$ independent on N_1. Therefore

$$\liminf_{N_1,N_2} \widehat{A}_{N_1,N_2,\tilde{\varepsilon}} \geqslant \max\left(\tilde{A}(q), A^1_{\tilde{\varepsilon}}(\eta;q), A^2_{\tilde{\varepsilon}}(\eta;q)\right), \tag{2.5}$$

with

$$A^1_{\tilde{\varepsilon}}(\eta;q) = \alpha(q+\eta)R^2 - \alpha\tilde{\varepsilon}\eta - \frac{\alpha}{2}\int \mathrm{d}\lambda \rho_{\mathrm{MP}}(\lambda;\alpha)\log(2(q+\eta)-\beta\lambda) - \alpha\log R - \frac{\alpha}{2}; \tag{2.6}$$

$$A^2_{\tilde{\varepsilon}}(\eta;q) = \alpha(q-\eta)R^2 - \alpha\tilde{\varepsilon}\eta - \frac{\alpha}{2}\int \mathrm{d}\lambda \rho_{\mathrm{MP}}(\lambda;\alpha)\log(2(q-\eta)-\beta\lambda) - \alpha\log R - \frac{\alpha}{2}. \tag{2.7}$$

Now we show that if $\bar{q}$ is the unique minimiser of $\tilde{A}(q)$ it is for η small enough

$$\tilde{A}(\bar{q}) = \max\left(\tilde{A}(\bar{q}), A^1_{\tilde{\varepsilon}}(\eta;\bar{q}), A^2_{\tilde{\varepsilon}}(\eta;\bar{q})\right),$$

which will conclude the proof.

To do so we introduce

$$\Delta_-(q;\eta) := \tilde{A}(q) - A^1_{\tilde{\varepsilon}}(\eta;q) = -\alpha\eta(R^2-\tilde{\varepsilon}) + \frac{\alpha}{2}\int \mathrm{d}\lambda \rho_{\mathrm{MP}}(\lambda;\alpha)\log\left(\frac{2(q+\eta)-\beta\lambda}{2q-\beta\lambda}\right); \tag{2.8}$$

$$\Delta_+(q;\eta) := \tilde{A}(q) - A^2_{\tilde{\varepsilon}}(\eta;q) = \alpha\eta(R^2+\tilde{\varepsilon}) + \frac{\alpha}{2}\int \mathrm{d}\lambda \rho_{\mathrm{MP}}(\lambda;\alpha)\log\left(\frac{2(q-\eta)-\beta\lambda}{2q-\beta\lambda}\right). \tag{2.9}$$

As a function of η, $\Delta_\pm(q;\eta)$ are continuous and differentiable, vanishing in $\eta = 0$ and with $\lim_{\eta\to+\infty}\Delta_\pm(q,\eta) = \pm\infty$. Moreover $\Delta_+(q;\eta)$ is uniformly convex and $\Delta_-(q;\eta)$ uniformly concave. Thus $\Delta_-(q;\eta)$ assumes a positive maximum iff the derivative in $\eta = 0$ is positive, that is

$$0 < -(R^2-\tilde{\varepsilon}) + \int \mathrm{d}\lambda \frac{\rho_{\mathrm{MP}}(\lambda;\alpha)}{2q-\beta\lambda} = \tilde{\varepsilon} - \partial_q\tilde{A}(q). \tag{2.10}$$

Likewise for $\Delta_+(q;\eta)$ is always positive iff $\partial_\eta\Delta_+(q;\eta)|_{\eta=0} \geqslant 0$, i.e.

$$0 \leqslant (R^2+\tilde{\varepsilon}) + \int \mathrm{d}\lambda \frac{\rho_{\mathrm{MP}}(\lambda;\alpha)}{2q-\beta\lambda} = \tilde{\varepsilon} + \partial_q\tilde{A}(q). \tag{2.11}$$

Combining (2.10) and (2.11) we get

$$-\tilde{\varepsilon} \leqslant \partial_q\tilde{A}(q)\big|_{q=\tilde{q}} < \tilde{\varepsilon},$$

that is $\tilde{q}$ is the unique stationary point of $\tilde{A}(q)$. With this choice of q, relation (2.5) gives

$$\liminf_{N_1,N_2} \widehat{A}_{N_1,N_2,\tilde{\varepsilon}} \geqslant \min_{q\geqslant\beta\lambda_+} \tilde{A}(q). \tag{2.12}$$

As $\tilde{\varepsilon}$ can be taken arbitrarily small, we recover (2.1). □

Now we pass to compute the free energy of $\mathrm{RBM}_{\sigma\sigma}$ by adapting the same method as before. Until the end of this section, Hamiltonian, partition function etc will be referred to the $\mathrm{RBM}_{\sigma\sigma}$. First step is to use the singular value decomposition of the matrix $\{\xi_{ij}\}$ to write

$$\frac{1}{\sqrt{N_1+N_2}}(x,\xi y)=\sqrt{1-\alpha}\sum_{i\in[N_1]}\sigma_i\tilde{x}_i\tilde{y}_i, \tag{2.13}$$

where $\tilde{x},\tilde{y}$ are related respectively to x,y by an orthogonal transformation. Note that this decomposition removes automatically N_2-N_1 cyclic coordinates of the second layer.

First of all we show that the part of the spectrum which falls away the support of the Marchenko–Pastur law is negligible for the free energy. The proof is adapted from [10] and we will stress only the most salient points of it.

Lemma 2.5. *There exists $c,C>0$ so that*

$$E[A_{N_1,N_2}1_{\{\max_{i\in[N_1]}\sigma_i>\bar{\sigma}\}}]\leqslant Ce^{-cN_1},$$

where $\bar{\sigma}:=1+\sqrt{\frac{\alpha}{1+\alpha}}$.

Proof. For any $a>\bar{\sigma}$ we compute

$$E[e^{\alpha\beta\sum_{i\in[N_1]}\sigma_i\tilde{x}_i\tilde{y}_i}]=E[e^{\beta\sum_{i\in[N_1]}\sigma_i\tilde{x}_i\tilde{y}_i}1_{\{\max_{i\in[N_1]}\sigma_i\leqslant a\}}]+E[e^{\beta\sum_{i\in[N_1]}\sigma_i\tilde{x}_i\tilde{y}_i}1_{\{\max_{i\in[N_1]}\sigma_i>a\}}].$$

The first summand is easily estimated by

$$E[e^{\beta\sum_{i\in[N_1]}\sigma_i\tilde{x}_i\tilde{y}_i}1_{\{\max_{i\in[N_1]}\sigma_i\leqslant a\}}]\leqslant E[e^{\beta a\sum_{i\in[N_1]}\tilde{x}_i\tilde{y}_i}]\leqslant E[e^{\beta a\|\tilde{x}\|\|\tilde{y}\|}]. \tag{2.14}$$

Arguing as in the proof of [10, proposition 1] we can write for a $C>0$

$$E[e^{\beta\sum_{i\in[N_1]}\sigma_i\tilde{x}_i\tilde{y}_i}1_{\{\max_{i\in[N_1]}\sigma_i>a\}}]\leqslant Ca\sqrt{2\pi(N_1+N_2)}e^{\frac{\beta^2}{2(N_1+N_2)}\|\tilde{x}\|^2\|\tilde{y}\|^2}. \tag{2.15}$$

Then (2.14) and (2.15) give an annealed bound for the free energy:

$$\limsup_{N_1,N_2}A_N\leqslant\max\left(\bar{\sigma}\beta\sqrt{\alpha(1-\alpha)}R^2,\ \frac{1}{2}\beta^2\alpha(1-\alpha)R_1^2R_2^2\right)<\infty. \tag{2.16}$$

With this bound at hand, we repeat *mutatis mutandis* the steps of the proof of [10, lemma 1] to get

$$E[A_{N_1,N_2}1_{\{\max_{i\in[N_1]}\sigma_i>\bar{\sigma}\}}]\leqslant\sqrt{\alpha(1-\alpha)\beta^2R^4+a^2}\sqrt{P(\|\Xi\Xi^T/N_2-\mathbb{I}\|_{op}\geqslant t+\sqrt{\lambda_+}-1)} \tag{2.17}$$

and the assert follows from (1.2). □

Proof of theorem 2.2. Now start from (2.13) and write for any a,b such that $ab>\beta^2(1-\alpha)\lambda_+$

$$\begin{aligned}Z_{N_1,N_2}&=\int\sigma_1(\mathrm{d}x)\sigma_2(\mathrm{d}y)\exp\left(\sqrt{1-\alpha}\beta\sum_{i\in[N_1]}x_iy_i\sigma_i-\frac{a}{2}\|x\|^2-\frac{b}{2}\|y\|^2+\frac{a}{2}R^2N_1+\frac{b}{2}R^2N_2\right)\\&=\exp\left(\frac{R^2}{2}(aN_1+bN_2)\right)\int\sigma_1(\mathrm{d}x)\sigma_2(\mathrm{d}y)\exp\left(-\frac{1}{2}\sum_{i=1}^{N_2}(z_i^T,M_iz_i)\right),\end{aligned} \tag{2.18}$$

where we have defined

$$z_i^T := \begin{cases} (x_i, y_i) & i = 1, \dots, N_1, \\ (0, y_i) & i = N_1 + 1, \dots, N_2 \end{cases}$$

and the 2×2 symmetric positively defined matrix

$$M_i := \begin{pmatrix} a & -\beta\sqrt{1-\alpha}\sigma_i \\ -\beta\sqrt{1-\alpha}\sigma_i & b \end{pmatrix}, \quad i \in [N_1] \qquad M_i := \mathrm{diag}(0, b), \quad i = N_1 + 1, \dots, N_2.$$

Thus by (1.9)

$$\begin{aligned} \varepsilon^2 \hat{Z}_{N_1,N_2,\varepsilon} &\leqslant \exp\left(\frac{R^2}{2}(aN_1 + bN_2)\right) \frac{(2\pi)^{\frac{N_1}{2}}(2\pi)^{\frac{N_2}{2}}}{S_{N_1}^R S_{N_2}^R} \int \frac{\mathrm{d}x\mathrm{d}y}{\sqrt{2\pi}^N} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_2}(z_i^T, M_i z_i)\right), \\ &= \exp\left(\frac{R^2}{2}(aN_1 + bN_2)\right) \frac{(2\pi)^{\frac{N_1}{2}}(2\pi)^{\frac{N_2}{2}}}{S_{N_1}^R S_{N_2}^R} b^{-\frac{N_2-N_1}{2}} \prod_{i=1}^{N_1}(ab - \beta^2(1-\alpha)\sigma_i^2)^{-\frac{1}{2}}. \end{aligned} \tag{2.19}$$

We introduce

$$\bar{A}(a,b) := \frac{R^2}{2}(a\alpha + b(1-\alpha)) - \log R - \frac{1}{2} - \frac{1-2\alpha}{2}\log b - \frac{\alpha}{2}\int \rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\log(ab - (1-\alpha)\beta^2\lambda) \tag{2.20}$$

and

$$\limsup_{N_1,N_2} A_{N_1,N_2} \leqslant \bar{A}(a,b), \quad \forall a, b > 0 \ : \ ab > (1-\alpha)\beta^2\lambda_+.$$

A direct inspection shows that $A(a,b)$ is jointly uniformly convex for any $\alpha \in [0, 1/2]$ if $ab > (1-\alpha)\beta^2\lambda_+$, therefore

$$\limsup_{N_1,N_2} A_{N_1,N_2} \leqslant \min_{ab>(1-\alpha)\beta^2\lambda_+} \bar{A}(a,b). \tag{2.21}$$

We record for later use the gradient coordinates

$$\partial_a \bar{A}(a,b) = \frac{\alpha R^2}{2} - \frac{\alpha}{2}\int \rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\frac{b}{ab - (1-\alpha)\beta^2\lambda}; \tag{2.22}$$

$$\partial_b \bar{A}(a,b) = \frac{(1-\alpha)R^2}{2} - \frac{1-2\alpha}{2b} - \frac{\alpha}{2}\int \rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\frac{a}{ab - (1-\alpha)\beta^2\lambda}. \tag{2.23}$$

For the reverse bound consider again spherical shells around S_{R,N_1} and S_{R,N_2} of thickness ε, that we name respectively $S_{1,\varepsilon}$ and $S_{2,\varepsilon}$. We split $S_{1,\varepsilon} = \mathbb{R}^{N_1} \setminus S_{1,\varepsilon}^c$ and $S_{2,\varepsilon} = \mathbb{R}^{N_2} \setminus S_{2,\varepsilon}^c$ and set $S_\varepsilon := S_{1,\varepsilon} \times S_{2,\varepsilon}$. We have

$$\varepsilon^2\widehat{Z}_{N_1,N_2,\varepsilon} := \varepsilon^2 \int \sigma_{N_1,R,\varepsilon}(\mathrm{d}x)\sigma_{N_2,R,\varepsilon}(\mathrm{d}y)\mathrm{e}^{-\beta H_N(x,y)}$$
$$= \exp\left(\frac{R^2}{2}(aN_1+bN_2)\right)\int_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}}\frac{\mathrm{d}x\mathrm{d}y}{|S_1||S_2|}\exp\left(-\frac{1}{2}\sum_{i=1}^{N_2}(z_i^T,M_iz_i)\right)$$
$$-\exp\left(\frac{R^2}{2}(aN_1+bN_2)\right)\int_{S_\varepsilon^c}\frac{\mathrm{d}x\mathrm{d}y}{|S_1||S_2|}\exp\left(-\frac{1}{2}\sum_{i=1}^{N_2}(z_i^T,M_iz_i)\right), \tag{2.24}$$

where we used again the representation (2.18). The free energy associated to the first summand was already computed above in the thermodynamic limit. Therefore we have to upper bound the second summand. We consider four contributions according to the following decomposition. Let $\kappa\in\{-1,1\}$ and put

$$S_\varepsilon^{\kappa,1} := \{x\in\mathbb{R}^{N_1}, y\in\mathbb{R}^{N_2} \;:\; \kappa(\|x\|^2-R^2N_1)\geqslant \varepsilon N_1\},$$
$$S_\varepsilon^{\kappa,2} := \{x\in\mathbb{R}^{N_1}, y\in\mathbb{R}^{N_2} \;:\; \kappa(\|y\|^2-R^2N_2)\geqslant \varepsilon N_2\}.$$

Thus $S_\varepsilon^c = \bigcup_{\kappa\in\{-1,1\},j\in\{1,2\}} S_\varepsilon^{\kappa,j}$. Moreover we pick $\eta>0$ small enough and set

$$M_i^{(\kappa,j)}(\eta) := \begin{pmatrix} a+(2-j)\kappa\eta & -\beta\sqrt{1-\alpha}\sigma_i \\ -\beta\sqrt{1-\alpha}\sigma_i & b+(j-1)\kappa\eta\end{pmatrix}, \quad i\in[N_1]$$
$$M_i^{(\kappa,j)}(\eta) := \mathrm{diag}(0, b+(j-1)\kappa\eta), \quad i=N_1+1,\dots,N_2.$$

Also, we put

$$Z_{N_1,N_2,\varepsilon}^{\kappa,j} := \exp\left(\frac{R^2}{2}(aN_1+bN_2)+\frac{(2-j)N_1\kappa\eta a}{2}+\frac{(j-1)N_2\kappa\eta b}{2}-\eta\varepsilon N_j\right)$$
$$\int_{S_\varepsilon^{\kappa,j}}\frac{\mathrm{d}x\mathrm{d}y}{|S_1||S_2|}\exp\left(-\frac{1}{2}\sum_{i=1}^{N_2}(z_i^T,M_i^{(\kappa,j)}(\eta)z_i)\right) \tag{2.25}$$

so that

$$(2.24) = -\sum_{\kappa\in\{-1,1\},j=1,2} Z_{N_1,N_2,\varepsilon}^{\kappa,j}.$$

We conclude that for any $\eta>0$ sufficiently small and a,b with $ab>(1-\alpha)\beta^2\lambda_+$

$$\limsup_{N_1,N_2}\widehat{A}_{N_1,N_2,\varepsilon}\geqslant \max(\bar{A}(a,b),\{A_\varepsilon^{\kappa,j}(\eta;a,b)\}_{\kappa\in\{-1,1\},j\in\{1,2\}}), \tag{2.26}$$

where

$$A_\varepsilon^{\kappa,j}(\eta;a,b) := -\eta\varepsilon+\frac{1}{2}\alpha a(R^2+(2-j)\kappa\eta)+\frac{1}{2}(1-\alpha)b(R^2+(j-1)\kappa\eta)$$
$$-\log R-\frac{1}{2}-\frac{1-2\alpha}{2}\log(b+(j-1)\kappa\eta)$$
$$-\frac{\alpha}{2}\int\rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\log((a+(2-j)\kappa\eta)(b+(j-1)\kappa\eta)-(1-\alpha)\beta^2\lambda).$$

In analogy with the proof of theorem 2.1 we define

$$\begin{aligned}\Delta_{\varepsilon}^{\kappa,j}(\eta) &:= \bar{A}(a,b) - A_{\varepsilon}^{\kappa,j}(\eta;a,b)\\ &= \eta\varepsilon - (2-j)\kappa\frac{\eta\alpha a}{2} - (j-1)\kappa\frac{\eta(1-\alpha)b}{2} + \frac{1-2\alpha}{2}\log(b+(j-1)\kappa\eta)\\ &\quad + \frac{\alpha}{2}\int \rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\log\left(\frac{(a+(2-j)\kappa\eta)(b+(j-1)\kappa\eta)-(1-\alpha)\beta^2\lambda}{ab-(1-\alpha)\beta^2\lambda}\right).\end{aligned}$$

We need to show that $\Delta_{\varepsilon}^{\kappa,j}(\eta) \geqslant 0$ for $\eta > 0$ small. As before, to do so it suffices to prove the derivative in the origin to be non-negative uniformly in $\varepsilon > 0$. Bearing in mind (2.22) and (2.23) we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\eta}\Delta_{\varepsilon}^{\kappa,j}(\eta)\big|_{\eta=0} &= \varepsilon + (2-j)\kappa\left(-\frac{a\alpha R^2}{2} + \frac{\alpha}{2}\int\rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\left(\frac{b}{ab-(1-\alpha)\beta^2\lambda}\right)\right)\\ &\quad + (j-1)\kappa\left(-\frac{b(1-\alpha)R^2}{2} + \frac{1-2\alpha}{2b} + \frac{\alpha}{2}\int\rho_{\mathrm{MP}}(\mathrm{d}\lambda;\alpha)\left(\frac{a}{ab-(1-\alpha)\beta^2\lambda}\right)\right)\\ &= \varepsilon - (2-j)\kappa\partial_a\bar{A} - (j-1)\kappa\partial_b\bar{A} \geqslant 0.\end{aligned}$$

Since the inequality must hold for any $\varepsilon > 0$, $\kappa \in \{-1,1\}$ and $j \in \{1,2\}$, we have to pick $(\bar{a},\bar{b}) = \arg\min\bar{A}$. Therefore

$$\limsup_{N_1,N_2} A_{N_1,N_2,\varepsilon} \geqslant \max(\bar{A}(\bar{a},\bar{b}), \{A^{\varepsilon,\kappa_1,\kappa_2}(\eta;\bar{a},\bar{b})\}_{\kappa_1,\kappa_2\in\{-1,1\}}) = \min_{ab>\beta^2(1-\alpha)\lambda_+}\bar{A}(a,b) \quad \forall\varepsilon>0,$$

which combined with (2.21) proves the theorem. □

3. Legendre equivalences of priors

In this section we explain the Legendre equivalence of spherical models on general terms. First of all we prove some *a priori* estimates ensuring the boundedness of the free energy in the thermodynamic limit. This will be used to cut the tails of the Gaussian distributions of the prior.

The first quick remark is that combining theorem 2.1 and a marginalisation we have

Corollary 3.1. *Let A_N be the free energy of $RBM_{\sigma,\gamma}$. Then $\lim_N A_N$ exists P-a.s.*

Next we focus on the Hopfield model with Gaussian prior previously defined.

Lemma 3.2. *Let A_{N_1,N_2} be the free energy of HM_{ρ}. There is $f(\lambda_+,\beta)$ continuous and bounded for which*

$$\limsup_{N_1,N_2} E[A_{N_1,N_2}] \leqslant f(\lambda_+,\beta). \tag{3.1}$$

Proof. Let $a > \lambda_+$ and set for brevity $\lambda^* := \max_{i\in[N_1]}\lambda_i$. We write

$$\frac{1}{N_1+N_2}E[\log Z_{N_1,N_2}] = \sum_{k\geqslant 0} E\left[\frac{1}{N_1+N_2}\log Z_{N_1,N_2} \mid (k+1)a \geqslant \lambda^* > ka\right] P((k+1)a \geqslant \lambda^* > ka).$$

Inside the conditional expectation we can bound $H_N(z) \leqslant a\|z\|^2$. Therefore

$$E\left[\frac{1}{N_1+N_2}\log Z_{N_1,N_2} \mid (k+1)a \geqslant \lambda^* > ka\right] \leqslant \frac{1}{N_1+N_2}\log\int \gamma_{N_1,\theta}(\mathrm{d}x)\exp\left(\|x\|^2(\beta a(k+1)) - Nr\left(\frac{\|x\|}{\sqrt{N}}\right)\right)$$
$$\leqslant \frac{1}{N_1+N_2}\max_{R\geqslant 0}\left(R^2(\beta a(k+1)) - N_1 r\left(\frac{R}{\sqrt{N_1}}\right)\right)$$
$$= \max_{R\geqslant 0}\left(R^2(\beta a(k+1)) - r(R)\right).$$

On the other hand by assumption

$$P((k+1)a \geqslant \lambda^* > ka) \leqslant 2\mathrm{e}^{-ca^2k^2N_1}, \quad c>0. \tag{3.2}$$

In conclusion

$$E[A_{N_1,N_2}] \leqslant \sum_{k\geqslant 0}\mathrm{e}^{-ca^2k^2N_1}\max_{R\geqslant 0}\left(R^2(\beta a(k+1)) - r(R)\right) =: f(a,\beta), \tag{3.3}$$

a continuous bounded function. In particular the estimate holds also for $a \to \lambda_+$. □

The analogue statement for RBM_{ρ^2}:

Lemma 3.3. *Let A_{N_1,N_2} be the free energy of RBM_{ρ^2}. There is $f(\lambda_+,\beta)$ continuous and bounded for which*

$$\limsup_{N_1,N_2} E[A_{N_1,N_2}] \leqslant f(\lambda_+,\beta). \tag{3.4}$$

Proof. Same proof as before, noting

$$\beta(1-\alpha)\sum_i \sigma_i x_i y_i - \sqrt{N_1N_2}\,r\left(\frac{\|x\|}{\sqrt{N_1}},\frac{\|y\|}{\sqrt{N_2}}\right) \leqslant \beta(1-\alpha)a\|x\|\|y\| - \sqrt{N_1N_2}\,r\left(\frac{\|x\|}{\sqrt{N_1}}\right)$$
$$\leqslant \max_{R_1,R_2}\left(\beta(1-\alpha)aR_1R_2 - r(R_1,R_2)\right)$$

for any $a > \sigma_+$. □

The above results immediately allow us to achieve the following useful lemma.

Lemma 3.4. *Let $R,\delta>0$, $N\in\mathbb{N}$. It holds for some $C,c>0$*

$$\int_{\{\|x\|^2\geqslant R^2N_1^{1+\delta}\}}\rho_{N_1}(\mathrm{d}x)\mathrm{e}^{-\beta H_{N_1,N_2}(x)} \leqslant C\mathrm{e}^{-cN_1^{1+\delta}}, \tag{3.5}$$

$$\int_{\{\|x\|^2\geqslant R^2N_1^{1+\delta}\}}\gamma_{N_1,\theta}(\mathrm{d}x)\sigma_{R,N_2}(\mathrm{d}y)\mathrm{e}^{-\beta H_{N_1,N_2}(x,y)} \leqslant C\mathrm{e}^{-cN_1^{1+\delta}}, \tag{3.6}$$

$$\int_{\{\|x\|^2\geqslant R^2N_1^{1+\delta}\}\cup\{\|y\|^2\geqslant R^2N_2^{1+\delta}\}}\rho^2_{N_1,N_2}(\mathrm{d}x\mathrm{d}y)\mathrm{e}^{-\beta H_{N_1,N_2}(x,y)} \leqslant C\mathrm{e}^{-cN_1^{1+\delta}}. \tag{3.7}$$

Proof. We prove only (3.5)–(3.7) are similar. Let us write

$$\int_{\{\|x\|^2 \geqslant R^2 N_1^{1+\delta}\}} \rho(\mathrm{d}x) \mathrm{e}^{-\beta H_{N_1,N_2}(x)} \leqslant \mathrm{e}^{-\frac{R^2 N_1^{1+\delta}}{2\theta}} \int_{\{\|x\|^2 \geqslant R^2 N_1^{1+\delta}\}} \gamma_{N_1,2\theta}(\mathrm{d}x) \mathrm{e}^{-\beta H_{N_1,N_2}(x) - N_1 r\left(\frac{\|x\|}{\sqrt{N_1}}\right)}$$
$$\leqslant \mathrm{e}^{-\frac{R^2 N_1^{1+\delta}}{2\theta}} Z_{N_1,N_2}(2\theta) = \exp\left((N_1+N_2)A_{N_1,N_2}(2\theta) - \frac{R^2 N_1^{1+\delta}}{2\theta}\right).$$

Here we have emphasised the dependence on θ of partition function and free energy. Then (3.5) follows from lemma 3.3. □

Now we are ready to prove the Legendre equivalences of theorem 1.1. We shall prove only (1.14), (1.16) and (1.18); the dual relations (1.15), (1.17) and (1.19) then follow directly, as one can easily verify the inverse Legendre transformation to be also well defined and involutive.

We start by (i), where we deal with a single Gaussian prior. Let $\varepsilon > 0$, $\delta > 0$. From now on we will systematically omit the dependence on δ of the objects we will operate with. Let further $r < N_1^{\delta}/\varepsilon$, $R_0 := 0$, $R_{r+1} := N_1^{\delta}$, $\{R_i\}_{i=1,\dots,r} \subset [0, N_1^{\delta})$ with $|R_{i+1} - R_i| < 2\varepsilon$, and decompose $\mathbb{R}^{N_1} := \bigcup_{i=0}^{r} S_{N_1,\varepsilon}^{[i]} \cup T$, where

$$S_{N_1,\varepsilon}^{[i]} := \{z \in \mathbb{R}^{N_1} \;\; R_i\sqrt{N_1} \leqslant \|z\|_2 \leqslant R_{i+1}\sqrt{N_1}\}, \quad T := \{z \in \mathbb{R}^{N_1} \;\; \|z\|_2 \geqslant {N_1}^{\delta+\frac{1}{2}}\}.$$

Comparing with (1.8) one easily sees that the $S_{\varepsilon,N_1}^{[i]}$ are spherical shells. We denote by $\sigma_{N_1,\varepsilon}^{[i]}$ the uniform distributions on these shells. Then we have

$$Z_{N_1,N_1}^{\rho} = \sum_{i=0}^{r} Z_{N_1,N_2}^{\rho}[i] + \tilde{Z}_{N_1,N_2}^{\rho}, \tag{3.8}$$

where

$$Z_{N_1,N_2}^{\rho}[i] := \int_{S_{N_1,\varepsilon}^{[i]}} \rho_{N_1}(\mathrm{d}x) \mathrm{e}^{-\beta H_{N_1,N_2}(x)}, \quad \tilde{Z}_{N_1,N_2}^{\rho} := \int_{T} \rho_{N_1}(\mathrm{d}x) \mathrm{e}^{-\beta H_{N_1,N_2}(x)}.$$

The tail term $\tilde{Z}_{N_1,N_2}^{\rho}$ gives a negligible contribution thanks to lemma 3.4 and we will ignore it all the time. Thus by (3.8) we get

$$\max_{i\in[r]} Z_{N_1,N_2}^{\rho}[i] \leqslant Z_{N_1,N_2,\beta}^{\rho} \leqslant \frac{N^{\delta}}{\varepsilon} \max_{i\in[r]} Z_{N_1,N_2}^{\rho}[i].$$

Therefore setting

$$A_{N_1,N_2}[i] := \left(\frac{1}{N_1+N_2} \log Z_{N_1,N_2}^{\rho}[i]\right) \tag{3.9}$$

we have

$$\max_{i\in[r]} \left(A_{N_1,N_2}[i]\right) \leqslant \frac{1}{N_1+N_2} \log Z_{N_1,N_2}^{\rho} \leqslant \max_{i\in[r]} \left(A_{N_1,N_2}[i]\right) + \frac{\delta \log N_1 - \log \varepsilon}{N_1+N_2}. \tag{3.10}$$

So the free energy of HM_{ρ} is given by the limit of $\max_{i\in[r]} \left(A_{N_1,N_2}[i]\right)$, provided it exists.

We notice now that by continuity for any $x \in S_{\varepsilon,N_1}^{[i]}$

$$\frac{\|x\|^2}{2\theta} + N_1 r\left(\frac{\|x\|}{\sqrt{N_1}}\right) = \frac{\tilde{R}_i^2 N_1}{2\theta} + N_1 r\left(\tilde{R}_i\right) + O(\varepsilon) \tag{3.11}$$

where $\tilde{R}_i \in [R_i, R_{i+1}]$. Therefore

$$Z^{\rho}_{N_1,N_2,\beta}[i] = \mathrm{e}^{-\frac{\tilde{R}_i^2 N_1}{2\theta} - N_1 r(\tilde{R}_i) - O(\varepsilon)} \frac{|S^{[i]}_{N_1,\varepsilon}|}{\sqrt{2\pi\theta}^{N_1}} \int \sigma^{[i]}_{N_1,\varepsilon}(\mathrm{d}x)\mathrm{e}^{-\beta H_{N_1,N_2}(x)} \tag{3.12}$$

and

$$\begin{aligned} A_{N_1,N_2}[i] &= \frac{N_1}{N_1+N_2}\left(\frac{1}{N_1}\log\left(\frac{|S^{[i]}_{\varepsilon,N_1}|}{\sqrt{2\pi\theta}^{N_1}}\right) - \frac{\tilde{R}_i^2}{2\theta} - r\left(\tilde{R}_i\right)\right) + \widehat{A}^{\varepsilon}_{N_1,N_2}\left(\tilde{R}_i\right) + \frac{O(\varepsilon)}{N_1+N_2} \\ &=: \tilde{A}_{N_1,N_2,\varepsilon}\left(\tilde{R}_i\right) + \frac{O(\varepsilon)}{N_1+N_2}. \end{aligned}$$

Thus we see that the existence of the thermodynamic limit is ensured by lemma 2.3 and we have

$$\lim_{N_1,N_2}\max_{i\in[r]}\left(A_{N_1,N_2}[i]\right) = \sup_{R>0}\lim_{N_1,N_2}\left(\tilde{A}_{N_1,N_2,\varepsilon}(R)\right) =: \sup_{R>0}\tilde{A}_{\varepsilon}(R).$$

By the uniform concavity of $\tilde{A}_{\varepsilon}(R)$ we have

$$\lim_{\varepsilon\to 0}\sup_{R>0}\tilde{A}_{\varepsilon}(R) = \sup_{R>0}\lim_{\varepsilon\to 0}\tilde{A}_{\varepsilon}(R) = -\frac{\alpha}{2}\log\theta + \alpha\sup_{R>0}\left(\log R + \frac{1}{2} + \alpha^{-1}A_{\sigma}(R,\beta) - r(R) - \frac{R^2}{2\theta}\right)$$

and the proof of (1.14) is concluded.

Note that in this argument the regularising function r plays essentially no role. So it can be set to zero and repeat verbatim all the previous steps for the $\mathrm{RBM}_{\sigma,\gamma}$. This way we obtain (1.16).

Finally we turn to RBM_{ρ^2}. Here we have to slice up both Gaussian priors and the previous construction easily extends. We just sketch the argument, stressing only the points in which it differs from above. Let $\varepsilon > 0$, $\delta > 0$, $r < N_1^{\delta}/\varepsilon$, $r' < N_2^{\delta}/\varepsilon$, $R_0, R'_0 := 0$, $R_{r+1} := N_1^{\delta}$, $R'_{r'+1} := N_2^{\delta}$, $\{R_i\}_{i=1,\dots,r} \subset [0, N_1^{\delta})$ with $|R_{i+1} - R_i| < 2\varepsilon$, $\{R'_i\}_{i=1,\dots,r'} \subset [0, N_2^{\delta})$ with $|R'_{i+1} - R'_i| < 2\varepsilon$. Decompose $\mathbb{R}^{N_1} \times \mathbb{R}^{N_2} := \bigcup_{i\in[r],j\in[r']} S^{[i]}_{N_1,\varepsilon} \times S'^{[j]}_{N_2,\varepsilon} \cup T$, where

$$\begin{aligned} S^{[i]}_{N_1,\varepsilon} &:= \{z \in \mathbb{R}^{N_1} \;\; R_i\sqrt{N_1} \leqslant \|z\|_2 \leqslant R_{i+1}\sqrt{N_1}\}, \\ S'^{[i]}_{N_2,\varepsilon} &:= \{z \in \mathbb{R}^{N_2} \;\; R'_i\sqrt{N_2} \leqslant \|z\|_2 \leqslant R'_{i+1}\sqrt{N_2}\}, \\ T &:= \{z \in \mathbb{R}^{N_1} \;\; \|z\|_2 \geqslant {N_1}^{\delta+\frac{1}{2}}\} \cup \{z \in \mathbb{R}^{N_2} \;\; \|z\|_2 \geqslant {N_2}^{\delta+\frac{1}{2}}\}. \end{aligned}$$

Again T can be neglected due to lemma 3.4. We have to evaluate

$$\begin{aligned} Z^{\rho^2}_{N_1,N_2}[i,j] &:= \int_{S^{[i]}_{N_1,\varepsilon}\times S'^{[j]}_{N_2,\varepsilon}} \rho^2_{N_1N_2}(\mathrm{d}x\mathrm{d}y)\mathrm{e}^{-\beta H_{N_1,N_2}(x,y)}, \\ &= \frac{|S^{[i]}_{N_1,\varepsilon} \times S'^{[j]}_{N_2,\varepsilon}|}{\sqrt{2\pi\theta}^{N_1+N_2}}\mathrm{e}^{-\frac{\tilde{R}_i^2N_1 + \tilde{R'}_j^2N_2}{2\theta} - \sqrt{N_1N_2}r(\tilde{R}_i,\tilde{R'}_j) + O(\varepsilon)} \int_{S^{[i]}_{N_1,\varepsilon}\times S'^{[j]}_{N_1,\varepsilon}} \sigma^{[i]}_{N_1,\varepsilon}(\mathrm{d}x)\sigma^{[i]}_{N_1,\varepsilon}(\mathrm{d}y)\mathrm{e}^{-\beta H_{N_1,N_2}(x,y)}, \end{aligned}$$

where the $\tilde{R}_i$ and $\tilde{R'}_i$ are introduced as before (see (3.11)). Therefore

$$A_{N_1,N_2,\beta}[i,j] := \frac{1}{N_1+N_2}\log Z^{\rho^2}_{N_1,N_2,\beta}[i,j]$$
$$= \frac{1}{N_1+N_2}\log\left(\frac{|S^{[i]}_{N_1,\varepsilon}\times S'^{[j]}_{N_2,\varepsilon}|}{\sqrt{2\pi\theta}^{N_1+N_2}}\right)+\frac{O(\varepsilon)}{N_1+N_2}-\frac{\tilde{R}_i^2+R'^2_j}{2\theta}$$
$$-\frac{\sqrt{N_1N_2}}{N_1+N_2}r\left(\tilde{R}_i,\tilde{R'}_j\right)+\widehat{A}_{N_1,N_2,\varepsilon}(\tilde{R}_i,\tilde{R}_j),$$

where $\widehat{A}_{N_1,N_2,\varepsilon}(\tilde{R}_i,\tilde{R'}_j)$ is the free energy of the RBM whose priors are the spherical shell measures of centres $\tilde{R}_i,\tilde{R'}_j$. The argument to pass to the thermodynamic limit is then the same, so (1.18) is obtained.

Acknowledgments

DT acknowledges GNFM-INDAM 'Progetto per Giovani Ricercatori' for financial support.

ORCID iDs

Daniele Tantari https://orcid.org/0000-0001-9982-5720

References

[1] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl Acad. Sci. USA* **79** 2554–8
[2] Fahlman S E, Hinton G E and Sejnowski T J 1983 Massively parallel architectures for Al: NETL *Thistle, and Boltzmann Machines National Conf. on Artificial Intelligence* (AAAI)
[3] Smolensky P 1986 *Information Processing in Dynamical Systems* (Cambridge, MA: MIT Press)
[4] Hinton G 2002 Training products of experts by minimizing contrastive divergence *Neural Comput.* **14** 1771–800
[5] Tieleman T 2008 Training restricted Boltzmann machines using approximations to the likelihood gradient *Proc. 25th Int. Conf. on Machine Learning* (ACM, APA)
[6] Hinton G E, Osindero S and Teh Y 2006 A fast learning algorithm for deep belief nets *Neural Comput.* **18** 1527–54
[7] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks *Science* **313** 504–7
[8] Salakhutdinov R and Hinton G 2009 Deep Boltzmann machines *Artificial Intelligence and Statistics (Proc. 12th Int. Conf. on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida)* vol 5 (Cambridge, MA: MIT Press)
[9] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)
[10] Genovese G and Tantari D 2015 Legendre duality of spherical and Gaussian spin glasses *Math. Phys. Anal. Geom.* **18** 1
[11] Bai Z and Silverstein J W 2010 *Spectral Analysis of Large Dimensional Random Matrices* (New York: Springer)
[12] Vershynin R 2012 *Introduction to the Non-Asymptotic Analysis of Random Matrices* (*Compressed Sensing*) (Cambridge: Cambridge University Press) pp 210–68
[13] Ben Arous G, Dembo A and Guionnet A 2001 Aging of spherical spin glasses *Probab. Theory Relat. Fields* **120** 1–67
[14] Barra A, Genovese G, Guerra F and Tantari D 2014 About a solvable mean field model of a Gaussian spin glass *J. Phys. A: Math. Theor.* **47** 155002

[15] Miolane L 2018 Phase transitions in spiked matrix estimation: information-theoretic analysis (arXiv:1806.04343)
[16] Hopfield J J 1984 Neurons with graded response have collective computational properties like those of two-state neurons *Proc. Natl Acad. Sci.* **81** 3088–92
[17] Barra A, Genovese G, Sollich P and Tantari D 2018 Phase diagram of restricted Boltzmann machines and generalised hopfield models with arbitrary priors *Phys. Rev.* E **97** 022310
[18] Bollè D, Nieuwenhuizen T M, Perez-Castillo I and Verbeiren T 2003 A spherical Hopfield model *J. Phys. A: Math. Gen.* **36** 10269
[19] Marks T K and Movellan J R 2001 Diffusion networks, products of experts, and factor analysis *Proc. Int. Conf. on Independent Component Analysis*
[20] Williams C K I and Agakov F V 2002 Products of Gaussians and probabilistic minor component analysis *Neural Comput.* **14** 1169–82
[21] Salakhutdinov R, Mnih A and Hinton G 2007 Restricted Boltzmann machines for collaborative filtering *Proc. 24th Int. Conf. on Machine Learning* (ACM)
[22] Hinton G E 2012 A practical guide to training restricted Boltzmann machines *Neural Networks: Tricks of the Trade* (Berlin: Springer) pp 599–619
[23] Williams C and Agakov F V 2002 An analysis of contrastive divergence learning in Gaussian Boltzmann machines *Institute for Adaptive and Neural Computation*
[24] Karakida R, Okada M and Amari S 2016 Dynamical analysis of contrastive divergence learning: restricted Boltzmann machines with Gaussian visible units *Neural Netw.* **79** 78–87
[25] Decelle C and Furtlehner A 2019 Gaussian-spherical restricted Boltzmann machines (arXiv:1910.14544)
[26] Barra A, Genovese G, Sollich P and Tantari D 2017 Phase transitions in restricted Boltzmann machines with generic priors *Phys. Rev.* E **96** 042156
[27] Auffinger A and Chen W-K 2014 Free energy and complexity of spherical bipartite models *J. Stat. Phys.* **157** 40–59
[28] Baik J and Lee J O 2017 Free energy of bipartite spherical Sherrington–Kirkpatrick model (arXiv:1711.06364)
[29] Barra A, Genovese G and Guerra F 2010 The replica symmetric approximation of the analogical neural network *J. Stat. Phys.* **140** 784
[30] Barra A, Genovese G and Guerra F 2011 Equilibrium statistical mechanics of bipartite spin systems *J. Phys. A: Math. Theor.* **44** 245002