Commun. Math. Phys.
Digital Object Identifier (DOI) https://doi.org/10.1007/s00220-021-04165-0

Communications in
Mathematical Physics

The Solution of the Deep Boltzmann Machine on the Nishimori Line

Diego Alberici[1], **Francesco Camilli**[2], **Pierluigi Contucci**[2], **Emanuele Mingione**[2]

[1] Communication Theory Laboratory, E.P.F.L., Lausanne, Switzerland
[2] Dipartimento di Matematica, Università di Bologna, Bologna, Italy,
E-mail: francesco.camilli2@unibo.it

Received: 10 February 2021 / Accepted: 1 July 2021

Abstract: The deep Boltzmann machine on the Nishimori line with a finite number of layers is exactly solved by a theorem that expresses its pressure through a finite dimensional variational problem of *min–max* type. In the absence of magnetic fields the order parameter is shown to exhibit a phase transition whose dependence on the geometry of the system is investigated.

1. Introduction

A deep (restricted) Boltzmann machine can be considered as a special case of the mean field multi-species spin glass model introduced in [11], further studied in [13,27]. Specifically the set of spins is arranged into a geometry made of consecutive layers and only interactions among spins belonging to adjacent layers are allowed. In particular intra-layer interactions are forbidden. Such architectural assumption makes it impossible to fulfill the positivity hypothesis under which the results of [11,27] were obtained. In fact the positivity property, encoded in an elliptic condition, requires dominant intra-group interaction with respect to inter-group ones. While the general deep (restricted) Boltzmann machine is still an unsolved problem (see nevertheless [1,3–5,12,18,23,24] for centered Gaussian interactions), we present here its exact and rigorous solution in a subregion of the phase space known as Nishimori line. In a previous paper [2] we have fully solved the elliptic multi-species model on the Nishimori line, where the property of replica symmetry, i.e. the concentration of the overlap, was shown to hold. Such property indeed is fully general on the Nishimori line, see [10] on this respect, and does not rely on any positivity assumption of the interactions. While the positivity properties carry with them the typical bounds of Guerra's method [19,20], here the technical support to control and solve the model is based on the presence, on the Nishimori line, of a set of identities relating magnetizations and overlaps expectations [16,25,26] and correlation inequalities [21]. Our work provides the first exact solution of a disordered Statistical

Published online: 27 July 2021

Mechanics model in a deep architecture and describes how the relative size of the layers influences the phase transition.

The relevance of the Nishimori line is twofold. On one side it provides the possibility to investigate the replica symmetric phase of the model through an exact solution for arbitrary strength of the interactions. On the other side it represents a bridge between a class of inference problems and Statistical Physics [26]. For instance the Sherrington–Kirkpatrick model on the Nishimori line corresponds to the Wigner Spiked model in the inference Bayesian optimal setting with binary signals [8,9]. Analogously, any multi-species mean-field model on the Nishimori line can be seen as a spatially coupled Wigner spiked model first introduced and studied in [6,7]. From the inference point of view here we deal with a deep spatially coupled Wigner spiked model with K layers, which in the case $K = 2$ coincides with the Wishart model (rank-one non-symmetric matrix estimation [9]).

The paper is organized as follows. In Section 2 we introduce the model and we present the main results in three theorems. Section 3 is a collection of tools and preliminary results, starting form the Nishimori identities and the correlation inequalities, up to the adaptive interpolation method. The proofs are contained in Section 4, Section 5 collects some conclusions and perspectives.

2. Definitions and Results

Consider a set of sites with cardinality N, divide it into K disjoint subsets, called *layers* and denoted by $\{L_r\}_{r=1,\dots,K}$ with cardinality $|L_r| = N_r$ and $\sum_{r=1}^{K} N_r = N$. To each site i we associate an Ising spin σ_i and we denote $\sigma = (\sigma_1, \dots, \sigma_N)$ a configuration of spins belonging to the space $\Sigma_N = \{+1, -1\}^N$. The Hamiltonian of the model is defined as:

$$H_N(\sigma) := -\sum_{r,s=1}^{K} \sum_{(i,j)\in L_r\times L_s} \tilde{J}_{ij}^{rs}\sigma_i\sigma_j - \sum_{r=1}^{K}\sum_{i\in L_r} \tilde{h}_i^r\sigma_i \tag{1}$$

where the interaction coefficients and the external fields are independent Gaussian random variables distributed as follows

$$\tilde{J}_{ij}^{rs} \overset{\text{iid}}{\sim} \mathcal{N}\left(\frac{\mu_{rs}}{2N}, \frac{\mu_{rs}}{2N}\right), \quad \tilde{h}_i^r \overset{\text{iid}}{\sim} \mathcal{N}(h_r, h_r)\,, \tag{2}$$

and the matrix $\mu := (\mu_{rs})_{r,s=1,\dots,K}$ and the vector $\mathbf{h} := (h_r)_{r=1,\dots,K}$ have non-negative entries. Furthermore μ has the following tridiagonal structure:

$$\mu = \begin{pmatrix} 0 & \mu_{12} & 0 & \cdots & 0 \\ \mu_{21} & 0 & \mu_{23} & \cdots & 0 \\ 0 & \mu_{32} & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \mu_{K-1,K} \\ 0 & 0 & 0 & \mu_{K,K-1} & 0 \end{pmatrix} \tag{3}$$

and is assumed to be symmetric without loss of generality. The geometrical architecture of the model is illustrated in Fig. 1.

We point out that the very special choice of the Gaussian distribution (2), having mean values and variances tied to be the same, is called *Nishimori line* in Physics literature

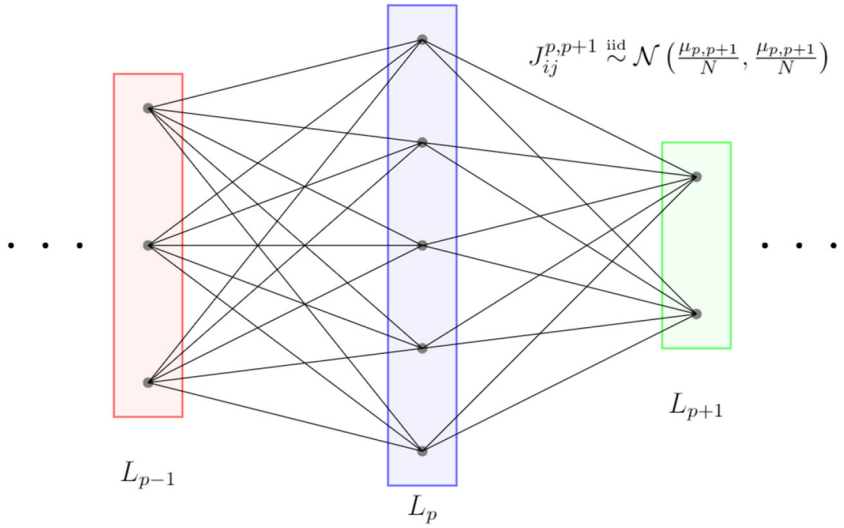

Fig. 1. Graph of the interactions between layers

[26]. We will recall correlation identities and inequalities holding on the Nishimori line in the next Section.

We denote

$$m_r(\sigma) := \frac{1}{N_r} \sum_{i \in L_r} \sigma_i \,, \qquad q_r(\sigma, \tau) := \frac{1}{N_r} \sum_{i \in L_r} \sigma_i \tau_i \,; \tag{4}$$

$$\mathbf{m}(\sigma) := (m_r(\sigma))_{r=1,\dots,K} \,, \quad \mathbf{q}(\sigma, \tau) := (q_r(\sigma, \tau))_{r=1,\dots,K} \tag{5}$$

with bold characters here and below standing for vectors and $\sigma, \tau \in \Sigma_N = \{-1, 1\}^N$. We also set

$$\Delta := (\alpha_r \mu_{rs} \alpha_s)_{r,s=1,\dots,K} \,, \quad \hat{\alpha} := \operatorname{diag}(\alpha_1, \alpha_2, \dots, \alpha_K) \,, \tag{6}$$

where $\alpha_r = N_r / N$ are called the *form factors*. Δ is the *effective interaction matrix* and encodes all the information on the interactions of the system. For later convenience we introduce also the matrix

$$M := (\mu_{rs} \alpha_s)_{r,s=1,\dots,K} \tag{7}$$

We notice that Δ and M are tridiagonal matrices too.

It is useful to express the Hamiltonian (1) in terms of centered Gaussian random variables plus a deterministic term (in vector notation):

$$H_N(\sigma) = -\frac{1}{\sqrt{2N}} \sum_{r,s=1}^{K} \sum_{(i,j) \in L_r \times L_s} J_{ij}^{rs} \sigma_i \sigma_j - \sum_{r=1}^{K} \sum_{i \in L_r} h_i^r \sigma_i - \frac{N}{2} (\mathbf{m}, \Delta \mathbf{m}) - N(\hat{\alpha} \mathbf{h}, \mathbf{m}) \tag{8}$$

where $(\cdot, \cdot)$ denotes the Euclidean inner product in $\mathbb{R}^K$ and

$$J_{ij}^{rs} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mu_{rs}) \,, \quad h_i^r \overset{\text{iid}}{\sim} \mathcal{N}(0, h_r) \,. \tag{9}$$

The random term in (8) corresponds to the Hamiltonian studied in [3], but the addition of a deterministic part changes the properties of the model.

We denote the random pressure per particle by

$$p_N := \frac{1}{N} \log \sum_{\sigma \in \Sigma_N} \exp\left(-H_N(\sigma)\right) \tag{10}$$

and its quenched average by

$$\bar{p}_N(\mu, \mathbf{h}) := \mathbb{E}\, p_N \,, \tag{11}$$

where $\mathbb{E}$ is the expectation with respect to all the Gaussian random variables.

Remark 1. While throughout this paper we keep the form factors α_r's constant as $N \to \infty$, all the results hold also under the weaker hypothesis that $N_r/N \to \alpha_r \in (0,1)$ (see also Remark 6, Sect. 4.3 for vanishing α_r's). Indeed any vanishing correction to α_r doesn't change the thermodynamic limit of the quenched pressure density (11). This can be seen proving by interpolation method that at given N the quenched pressure is a Lipschitz function of Δ w.r.t. the entrywise matrix norm $\sum_{r,s\le K} |\Delta_{r,s}|$.

The (random) Boltzmann–Gibbs average will be denoted by

$$\langle \cdot \rangle_N := \frac{\sum_{\sigma\in\Sigma_N} e^{-H_N(\sigma)}(\cdot)}{Z_N}, \quad Z_N := \sum_{\sigma\in\Sigma_N} e^{-H_N(\sigma)}. \tag{12}$$

To help the presentation we will occasionally make explicit the dependence of the Boltzmann–Gibbs measure on further parameters by using sub and superscripts, for instance $\langle\cdot\rangle_{N,t}^{(\epsilon)}$. In the previous definitions (10)–(12) we have chosen to reabsorb the inverse absolute temperature β in the parameters μ_{rs} and h_r. The first result of this paper is the computation of the random pressure (10) in the thermodynamic limit.

Theorem 1. (Solution of the model) *The random pressure* (10) *of a K-layer deep Boltzmann machine on the Nishimori line converges almost surely in the thermodynamic limit and its value is given by a K-dimensional variational principle:*

$$\lim_{N\to\infty} p_N \overset{\text{a.s.}}{=} \lim_{N\to\infty} \bar{p}_N(\mu, \mathbf{h}) = \sup_{\mathbf{x}_o} \inf_{\mathbf{x}_e} p_{var}(\mathbf{x}; \mu, \mathbf{h}) \,, \tag{13}$$

where $\mathbf{x}_o$ and $\mathbf{x}_e$ denote the vectors of the odd and even components of the order parameter $\mathbf{x} \in [0,1)^K$ respectively,

$$p_{var}(\mathbf{x}; \mu, \mathbf{h}) := \sum_{r=1}^{K} \alpha_r\, \psi\left((M\mathbf{x})_r + h_r\right) + \sum_{r=1}^{K} \frac{\Delta_{r,r+1}}{2} \left[(1-x_r)(1-x_{r+1}) - 2x_r x_{r+1}\right] \tag{14}$$

and for any $x \ge 0$

$$\psi(x) := \mathbb{E}_z \log 2\cosh\left(z\sqrt{x} + x\right) \,, \quad z \sim \mathcal{N}(0,1). \tag{15}$$

Moreover, defining $\bar{\mathbf{x}}$ as the solution of the variational problem (13)*, we have*

$$\lim_{N\to\infty} \mathbb{E}\langle q_r \rangle_N = \lim_{N\to\infty} \mathbb{E}\langle m_r \rangle_N = \bar{x}_r \tag{16}$$

for every $r = 1, \dots, K$ and for all the points of the phase space $(\mu, \hat{\alpha}, \mathbf{h})$ where $\bar{x}$ is $\mathbf{h}$-differentiable and the matrix Δ is invertible.

The proof of Theorem 1 relies on the adaptive interpolation method [8] combined with a concentration result and with the Nishimori identities, that will be presented in the next section. The main difference with the model solved in [2] is that the matrix Δ is not definite, indeed its eigenvalues have alternating signs. This entails that the remainder identified by interpolation has not a definite sign and cannot be discarded a priori at the expense of an inequality. Moreover, the concentration of the overlap strongly depends on a notion of regularity of the path followed by the adaptive interpolation. Hence one has to carefully choose a path that is regular and allows also to exploit the convexities of the two sums involved in the functional (14).

Secondly, we focus on the properties of the consistency equation obtained from the optimization problem (13) when the matrix Δ is invertible, that is when K is even. The stability of the optimizers of (13) is a more delicate problem with respect to the convex multi-species case [2], due to the min-max nature of the variational principle. In the following, given a square matrix A we denote by $\rho(A)$ its spectral radius and by $A^{(eo)}$ the submatrix of A obtained by keeping only even rows and odd columns of A. An analogous definition is given for $A^{(oe)}$, $A^{(oo)}$, $A^{(ee)}$. Notice that, when K is even, $\Delta^{(eo)}$ is an upper triangular $K/2 \times K/2$ square matrix with non-zero diagonal elements and therefore it is invertible. Similar considerations hold for the sub-matrix $\Delta^{(oe)} = [\Delta^{(eo)}]^T$. We prove the following

Theorem 2. *Let K be even and $\mathbf{h} = 0$. If $\rho([M^2]^{(oo)}) < 1$ then $\mathbf{x} = 0$ is the unique solution to the variational problem (13). Conversely, if $\rho([M^2]^{(oo)}) > 1$ then the solution of (13) is a vector $\mathbf{x} = \bar{\mathbf{x}}(M)$ with strictly positive components satisfying the consistency equation:*

$$x_r \;=\; \mathbb{E}_z \tanh\left(z\sqrt{(M\mathbf{x})_r} \;+\; (M\mathbf{x})_r\right) \quad \forall\, r = 1, \dots, K \tag{17}$$

where z denotes a standard Gaussian random variable.

The proof of Theorem 2 amounts to the computation of the Hessian matrix of an auxiliary function introduced later and in a check of its eigenvalues. The peculiar form of the consistency equations due to the structure (3) plays a central role. Theorem 2 implies the existence of a phase transition in our model localized at zero magnetic field and unitary spectral radius as discussed in Remark 2 below. The following Proposition further clarifies the structure of the phase transition and how the system's geometry, encoded in the form factors α_r's, can influence it.

Proposition 1. *For any given interaction matrix μ, we have*

$$\sup_{\alpha_1,\dots,\alpha_K} \rho\left([M^2]^{(oo)}\right) \;=\; \frac{1}{4}\max_r \mu^2_{r,r+1} \tag{18}$$

where the sup *on the l.h.s. is taken over the form factors $\alpha_1, \dots, \alpha_K \geq 0$, $\sum_{r=1}^K \alpha_r = 1$ and the* max *on the r.h.s. is taken over $r = 1, \dots, K-1$. Furthermore the* sup *on the l.h.s. of (18) is attained if and only if one of the following conditions is verified:*

(a) *there exists $r^* \in \{1, \dots, K-1\}$ such that*

$$\alpha_{r^*} \;=\; \alpha_{r^*+1} \;=\; \frac{1}{2} \quad , \quad \mu_{r^*,\,r^*+1} = \max_r \mu_{r,r+1}\;; \tag{19}$$

(b) *there exists* $r^* \in \{2, \dots, K-1\}$ *such that*

$$\alpha_{r^*} = \alpha_{r^*-1} + \alpha_{r^*+1} = \frac{1}{2} \quad , \quad \mu_{r^*-1,\, r^*} = \mu_{r^*,\, r^*+1} = \max_r \mu_{r,r+1} \, . \tag{20}$$

Remark 2. For even K, Proposition 1 together with Theorems 1 and 2 show that if the interaction strengths $\mu_{r,r+1} < 2$ for all $r = 1, \dots, K-1$, then the magnetisations and the overlaps vanish as $N \to \infty$ for every choice of the form factors $(\alpha_1, \dots, \alpha_K) \in (0,1)^K$. By Theorem 2 $\bar{\mathbf{x}}$ is not identically zero on the space of parameters $(\mu, \hat{\alpha})$, hence the limiting quenched pressure (13) cannot be an analytic function.

Proposition 1 also shows that as soon as $\mu_{r,r+1} > 2$ for some $r = 1, \dots, K-1$, then, by suitably localizing only two extensive layers near the maximal interaction (condition (19)), their magnetisations and overlaps turn out to be positive in the limit $N \to \infty$.

Finally, we prove a uniqueness result that holds for arbitrary spectral radius.

Theorem 3. *Let* $h_r > 0 \; \forall r = 1, \dots, K$ *. The consistency equation*

$$x_r = \mathbb{E}_z \tanh\left(z \sqrt{(M\mathbf{x})_r + h_r} + (M\mathbf{x})_r + h_r \right) \quad \forall r = 1, \dots, K \tag{21}$$

admits a unique solution $\mathbf{x} = \bar{\mathbf{x}}(M, \mathbf{h}) \in (0,1)^K$.

3. Preliminary Results

3.1. Nishimori identities and correlation inequalities. The main thermodynamic properties of the model are consequences of a family of identities and inequalities for correlation functions that are due to the specific setting (2). The identities were introduced in the original work by Nishimori [25] while the inequalities were proved in [21,22]. The proof of the Nishimori identities can be found in the book [15] (Paragraph 2.6). In particular, for our purposes we will use the following:

$$\mathbb{E}[\langle \sigma_i \rangle_N^{2n}] = \mathbb{E}[\langle \sigma_i \rangle_N^{2n-1}] \, , \quad n = 1, 2, 3, \dots \tag{22}$$

$$\mathbb{E}[\langle \sigma_i \sigma_j \rangle_N^{2}] = \mathbb{E}[\langle \sigma_i \sigma_j \rangle_N] \, . \tag{23}$$

From the previous relations it follows that on the Nishimori line magnetizations and overlaps moments coincide. This can be seen by

$$\mathbb{E}[\langle q_s \rangle_N] = \sum_{i \in L_s} \frac{1}{N_s} \mathbb{E}[\langle \sigma_i \rangle_N^2] = \sum_{i \in L_s} \frac{1}{N_s} \mathbb{E}[\langle \sigma_i \rangle_N] = \mathbb{E}[\langle m_s \rangle_N] \, , \tag{24}$$

$$\mathbb{E}[\langle q_r q_s \rangle_N] = \sum_{(i,j) \in L_r \times L_s} \frac{\mathbb{E}[\langle \sigma_i \sigma_j \rangle_N^2]}{N_r N_s} = \sum_{(i,j) \in L_r \times L_s} \frac{\mathbb{E}[\langle \sigma_i \sigma_j \rangle_N]}{N_r N_s} = \mathbb{E}[\langle m_r m_s \rangle_N] \, , \tag{25}$$

where the expectations $\langle q_s \rangle_N$ and $\langle q_r q_s \rangle_N$ are taken with respect to the replicated Gibbs measure. As a consequence we have:

$$\mathbb{E}\Big\langle (\mathbf{q}, \Delta \mathbf{q}) \Big\rangle_N = \mathbb{E}\Big\langle (\mathbf{m}, \Delta \mathbf{m}) \Big\rangle_N \, . \tag{26}$$

Concerning the correlation inequalities on the Nishimori line [16,21,22] (see also Theorem 2.18 in [15] for a straightforward proof) we have that:

$$\frac{\partial \bar{p}_N}{\partial h_r} = \frac{1}{2N} \sum_{i \in L_r} \mathbb{E}[1 + \langle \sigma_i \rangle_N] = \frac{\alpha_r}{2} \left(1 + \mathbb{E}\langle m_r \rangle_N\right) \geq 0 \,, \tag{27}$$

$$\frac{\partial^2 \bar{p}_N}{\partial h_r \partial h_s} = \frac{\alpha_r}{2} \frac{\partial \mathbb{E}\langle m_r \rangle_N}{\partial h_s} = \frac{1}{2N} \sum_{(i,j) \in L_r \times L_s} \mathbb{E}\left[\left(\langle \sigma_i \sigma_j \rangle_N - \langle \sigma_i \rangle_N \langle \sigma_j \rangle_N \right)^2 \right] \geq 0 \,. \tag{28}$$

Hence both the quenched pressure per particle and the magnetizations are non-decreasing with respect to each parameter h_r, $r = 1, \ldots, K$.

3.2. One-body system on the Nishimori line. It is useful to consider the following simple Hamiltonian on the Nishimori line, where only one-body interactions are taken into account:

$$H_N^{(0)}(\sigma) := -\sum_{i=1}^{N} (z_i \sqrt{h} + h)\, \sigma_i \,, \quad z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \tag{29}$$

with $h > 0$. It is easy to show that the pressure of this model coincides with the function $\psi(h)$ defined by (15):

$$p_N^{(0)} := \frac{1}{N} \mathbb{E} \log \sum_{\sigma \in \Sigma_N} e^{-H_N^{(0)}(\sigma)} = \psi(h) \,. \tag{30}$$

Since the Boltzmann–Gibbs average of a spin in the one body system equals $\langle \sigma_1 \rangle_N^{(0)} = \tanh\left(z_1 \sqrt{h} + h\right)$, the Nishimori identities entail the following identities:

$$\mathbb{E} \tanh^{2n-1}\left(z\sqrt{h} + h\right) = \mathbb{E} \tanh^{2n}\left(z\sqrt{h} + h\right) \tag{31}$$

for every $n \in \mathbb{N}$, $n \geq 1$. Starting from expression (15) we are going to determine the sign of the first derivatives of ψ. Gaussian integration by parts and identity (31) for $n = 1$ show that

$$\psi'(h) = \frac{1}{2}\left(1 + \mathbb{E} \tanh\left(z\sqrt{h} + h\right)\right) > 0 \,. \tag{32}$$

Using again Gaussian integration by parts and identity (31) for $n = 1, 2$, one finds:

$$\psi''(h) = \frac{1}{2} \mathbb{E}\left[\left(1 - \tanh^2\left(z\sqrt{h} + h\right)\right)^2 \right] > 0 \,. \tag{33}$$

The sign of the third derivative can be obtained avoiding Gaussian integration by parts. Indeed by setting $y = z\sqrt{h} + h$, replacing $\frac{z}{2\sqrt{h}} + 1 = \frac{y+h}{2h}$ in the computations and using the identities (31) for $n = 2, 3$, one finds:

$$\psi'''(h) = -\frac{1}{h} \mathbb{E}\left[\left(1 - \tanh^2 y\right)^2 y \tanh y \right] - \mathbb{E}\left[\left(1 - \tanh^2 y\right)^2 \tanh^2 y \right] < 0 \,. \tag{34}$$

The convexity of ψ will be crucial in the proof of Theorem 1. In particular, we will use the following

Lemma 1. *The function*

$$f(\mathbf{x}) := \sum_{r=1}^{K} \alpha_r \, \psi((M\mathbf{x})_r) \tag{35}$$

is convex for $\mathbf{x}$ *such that* $M\mathbf{x} \geq 0$ *component-wise.*

Proof. ψ is convex on $\mathbb{R}_{\geq 0}$ by equation (33). Then, using the linearity of $(M\mathbf{x})_r$, it is easy to verify that for any $\lambda \in [0, 1]$ and $\mathbf{x}_1, \mathbf{x}_2 \in A$ we have:

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \;\leq\; \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \; . \tag{36}$$

□

In the proof of Theorem 3 we will use the following

Lemma 2. *Let* z *be a standard Gaussian random variable. The function*

$$F(h) := \mathbb{E} \tanh\left(z\sqrt{h} + h\right) \tag{37}$$

is strictly positive, increasing and concave for $h > 0$.

Proof. It follows immediately by equations (31), (33), (34), since $F = 2\,\psi' - 1$. □

Remark 3. As a consequence the function F is invertible on $[0, \infty)$. Its inverse F^{-1} is non negative and increasing on $[0, 1)$. Moreover one has

$$\lim_{x \to 1^-} F^{-1}(x) = +\infty \, . \tag{38}$$

3.3. Interpolating model. We now introduce an interpolating model that compares the original model with a one-body model with suitably tuned external field.

Definition 1 (*Interpolating model*). Let $t \in [0, 1]$. The Hamiltonian of the interpolating model is:

$$\begin{aligned} H_\sigma(t) := &-\frac{\sqrt{1-t}}{\sqrt{2N}} \sum_{r,s=1}^{K} \sum_{(i,j) \in L_r \times L_s} J_{ij}^{rs} \sigma_i \sigma_j - (1-t)\,\frac{N}{2}\,(\mathbf{m}, \Delta \mathbf{m}) \; + \\ &- \sum_{r=1}^{K} \sum_{i \in L_r} \left(\sqrt{Q_{\epsilon,r}(t)}\, J_i^r + Q_{\epsilon,r}(t) \right) \sigma_i \; - \sum_{r=1}^{K} \sum_{i \in L_r} h_i^r \sigma_i - N(\hat{\alpha}\mathbf{h}, \mathbf{m}) \end{aligned} \tag{39}$$

with $J_i^r \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ independent of all the other Gaussian random variables, and

$$\mathbf{Q}_\epsilon(t) := \boldsymbol{\epsilon} + M \int_0^t \mathbf{q}_\epsilon(s)\, ds, \quad \epsilon_r \in [s_N, 2s_N], \; s_N \propto N^{-\frac{1}{16K}} \; .$$

Here $\mathbf{Q}_\epsilon =: (Q_{\epsilon,r})_{r=1,\dots,K}$, while $\mathbf{q}_\epsilon := (q_{\epsilon,r})_{r=1,\dots,K}$ denotes a vector of K non-negative functions that will be suitably chosen in the following.

Now we can write the sum rule, which is contained in the following proposition.

Proposition 2 (Sum rule). *The quenched pressure of the model rewrites as:*

$$\bar{p}_N(\mu, \mathbf{h}) = \mathcal{O}(s_N) + \sum_{r=1}^{K} \alpha_r \, \psi(Q_{\epsilon,r}(1) + h_r) + \int_0^1 dt \left[\frac{(\mathbf{1} - \mathbf{q}_\epsilon(t), \Delta(\mathbf{1} - \mathbf{q}_\epsilon(t)))}{4} - \frac{(\mathbf{q}_\epsilon(t), \Delta \mathbf{q}_\epsilon(t))}{2} \right] + \frac{1}{4} \int_0^1 dt \, R_\epsilon(t, \mu, \mathbf{h}) \,, \tag{40}$$

where the remainder is:

$$R_\epsilon(t, \mu, \mathbf{h}) = \mathbb{E} \Big\langle (\mathbf{m} - \mathbf{q}_\epsilon(t), \Delta(\mathbf{m} - \mathbf{q}_\epsilon(t))) \Big\rangle_{N,t}^{(\epsilon)} \,. \tag{41}$$

Proof. We stress that the interpolating model is on the Nishimori line for any $t \in [0, 1]$, as can be seen by direct inspection. This allows us to use the identities and inequalities for any value of the interpolating parameter. See Proof of Proposition 2 in [2] for the details. □

The tridiagonal form of Δ allows us to specialize the previous sum rule as follows:

$$\begin{aligned} \bar{p}_N(\mu, \mathbf{h}) \; &= \mathcal{O}(s_N) + \sum_{r=1}^{K} \alpha_r \, \psi \left(Q_{\epsilon,r}(1) + h_r \right) \\ &+ \sum_{r=1}^{K} \frac{\Delta_{r,r+1}}{2} \int_0^1 dt \, \left[(1 - q_{\epsilon,r}(t))(1 - q_{\epsilon,r+1}(t)) - 2 q_{\epsilon,r}(t) q_{\epsilon,r+1}(t) \right] \\ &+ \sum_{r=1}^{K} \frac{\Delta_{r,r+1}}{2} \int_0^1 dt \, \mathbb{E} \big\langle (m_r - q_{\epsilon,r}(t)) \, (m_{r+1} - q_{\epsilon,r+1}(t)) \big\rangle_{N,t}^{(\epsilon)} \,, \end{aligned} \tag{42}$$

or better, using the notation introduced for Theorem 2,

$$\begin{aligned} \bar{p}_N(\mu, \mathbf{h}) \; &= \mathcal{O}(s_N) + \sum_{r=1}^{K} \alpha_r \, \psi \left(Q_{\epsilon,r}(1) + h_r \right) \\ &+ \frac{1}{2} \int_0^1 dt \, \left[(\mathbf{1}_o - \mathbf{q}_{\epsilon,o}(t), \Delta^{(oe)}(\mathbf{1}_e - \mathbf{q}_{\epsilon,e}(t))) - 2(\mathbf{q}_{\epsilon,o}(t), \Delta^{(oe)} \mathbf{q}_{\epsilon,e}(t)) \right] \\ &+ \frac{1}{2} \int_0^1 dt \, \mathbb{E} \big\langle (\mathbf{m}_o - \mathbf{q}_{\epsilon,o}(t), \Delta^{(oe)}(\mathbf{m}_e - \mathbf{q}_{\epsilon,e}(t))) \big\rangle_{N,t}^{(\epsilon)} \,, \end{aligned} \tag{43}$$

where again the subscripts o, e denote the odd or even components of a vector, $\mathbf{1} := (1)_{r=1,\dots,K}$. We also denote

$$\mathbf{Q}_{\epsilon,o}(t) = \boldsymbol{\epsilon}_o + M^{(oe)} \int_0^t \mathbf{q}_{\epsilon,e}(s) \, ds \,, \quad \mathbf{Q}_{\epsilon,e}(t) = \boldsymbol{\epsilon}_e + M^{(eo)} \int_0^t \mathbf{q}_{\epsilon,o}(s) \, ds \,. \tag{44}$$

The sum rules (42), (43) motivate the definition of the variational pressure (14) that for future convenience can be rewritten as:

$$p_{var}(\mathbf{x};\mu,\mathbf{h}) = \sum_{r=1}^{K} \alpha_r \, \psi\left((M\mathbf{x})_r + h_r\right) + \frac{(\mathbf{1}_o - \mathbf{x}_o, \Delta^{(oe)}(\mathbf{1}_e - \mathbf{x}_e))}{2} - (\mathbf{x}_o, \Delta^{(oe)}\mathbf{x}_e)\,. \tag{45}$$

Remark 4. The variational function p_{var} is convex in the even components $\mathbf{x}_e$ and the odd components $\mathbf{x}_o$ separately. This is due to the fact that the two bilinear forms in (45) have vanishing second derivatives w.r.t. pure odd or even components, while the terms containing ψ are convex by Lemma 1.

The sum rule exhibits a remainder (namely (41)) to deal with. Let us first introduce the following

Definition 2 (*Regularity of* $\boldsymbol{\epsilon} \longmapsto \mathbf{Q}_\epsilon(\cdot)$). We will say that the map $\boldsymbol{\epsilon} \longmapsto \mathbf{Q}_\epsilon(\cdot)$ is *regular* if

$$\det\left(\frac{\partial \mathbf{Q}_\epsilon(t)}{\partial \boldsymbol{\epsilon}}\right) \geq 1 \quad \forall t \in [0,1] \tag{46}$$

This has to be combined with Liouville's formula, a standard analysis result that we report here for the reader's convenience.

Lemma 3 (Liouville's formula). *Consider two matrices whose elements depend on a real parameter:* $\Phi(t)$, $A(t)$. *Suppose that* Φ *satisfies the Cauchy problem*

$$\begin{cases} \dot{\Phi}(t) = A(t)\,\Phi(t) \\ \Phi(0) = \Phi_0 \end{cases} . \tag{47}$$

Then:

$$\det(\Phi(t)) = \det(\Phi_0)\, \exp\left\{\int_0^t ds\, \mathrm{Tr}(A(s))\right\} \tag{48}$$

Now, the remainder (41) can be proved to concentrate under the regularity hypothesis, as stated in the following

Lemma 4 (Concentration). *Suppose* $\boldsymbol{\epsilon} \longmapsto \mathbf{Q}_\epsilon(\cdot)$ *is a regular map. For every* $r = 1, \dots, K$ *consider the quantity*

$$\mathcal{L}_r := \frac{1}{N_r} \sum_{i \in L_r} \left(\sigma_i + \frac{J_i^r \sigma_i}{2\sqrt{Q_{\epsilon,r}(t)}}\right), \quad J_i^r \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \tag{49}$$

and introduce the $\boldsymbol{\epsilon}$*-average:*

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\cdot] = \prod_{r=1}^{K} \left(\frac{1}{s_N} \int_{s_N}^{2s_N} d\epsilon_r\right)(\cdot)\,. \tag{50}$$

We have:

$$\mathbb{E}_{\boldsymbol{\epsilon}} \mathbb{E}\left\langle \left(\mathcal{L}_r - \mathbb{E}\langle \mathcal{L}_r \rangle_{N,t}^{(\epsilon)}\right)^2 \right\rangle_{N,t}^{(\epsilon)} \longrightarrow 0 \quad as\ N \to \infty \tag{51}$$

and

$$\mathbb{E}\Big\langle \Big(m_r - \mathbb{E}\langle m_r\rangle_{N,t}^{(\epsilon)}\Big)^2\Big\rangle_{N,t}^{(\epsilon)} \le 4\,\mathbb{E}\Big\langle \Big(\mathcal{L}_r - \mathbb{E}\langle \mathcal{L}_r\rangle_{N,t}^{(\epsilon)}\Big)^2\Big\rangle_{N,t}^{(\epsilon)} \tag{52}$$

for every $r = 1, \dots, K$ *. Therefore the magnetization (or the overlap) concentrates in* $\boldsymbol{\epsilon}$*-average.*

The proof is carried out by treating the thermal and disordered fluctuations of $\mathcal{L}_r$ separately. Actually, an estimate on the L^2-convergence speed of the random pressure to the quenched one as $N \to \infty$ is required and it will be given in the proofs section below. See Lemma 3 and Appendix A in [2] for the details. The role of $\boldsymbol{\epsilon}$ is that of a regularizing perturbation and it is crucial for the proof. Its introduction intuitively allows to avoid critical points where the limiting pressure presents singularities and concentration may not occur, thus helping us to select always the stable state of the system. Indeed, for vanishing external magnetic fields $\mathbf{h} = 0$ and in absence of $\boldsymbol{\epsilon}$, the system remains stuck in a vanishing average magnetization state because of the resulting spin flip symmetry in the Hamiltonian. However, as stated in Theorem 2 in the appropriate range of parameters the latter is thermodynamically unstable, meaning that any arbitrarily small magnetic field would bring the magnetization to positive values.

4. Proofs

4.1. Proof of Theorem 1. The almost sure equality in (13) is a standard result based on the following concentration inequality:

Proposition 3. *There exists* $C = C(\mu, \mathbf{h}) > 0$ *such that for every* $x > 0$

$$\mathbb{P}\left(|p_N - \bar{p}_N(\mu, \mathbf{h})| \ge x\right) \le 2\exp\left(-\frac{Nx^2}{4C}\right). \tag{53}$$

As a consequence

$$\mathbb{E}[(p_N - \bar{p}_N(\mu, \mathbf{h}))^2] \le \frac{8C}{N}. \tag{54}$$

Proof. The random pressure p_N is a Lipschitz function of the independent standard Gaussian variables $\hat{J} = (J_{ij}^{rs}/\sqrt{\mu_{rs}})_{i,j,r,s}$, $\hat{h} = (h_i^r/\sqrt{h_r})_{i,r}$. Indeed:

$$N^2\,|\nabla_{\hat{J},\hat{h}}\, p_N|^2 \le N\left(\frac{(\mathbf{1}, \Delta\mathbf{1})}{2} + (\hat{\alpha}\mathbf{h}, \mathbf{1})\right) \equiv CN \tag{55}$$

The inequality (53) then follows by a known concentration property of the Gaussian measure (see Theorem 1.3.4 in [29]). A tail integration finally leads to (54). □

Since the r.h.s. in (53) is summable the Borel-Cantellli Lemma guarantees almost sure convergence. Now we move to the proof of the variational principle, i.e. the second equality in (13) which is going to be achieved through upper and lower bounds. For what follows, we neglect all the sub and superscripts in the Boltzmann–Gibbs averages, except for the t-dependence.

Lower bound. We select a path contained in $[0,1)^K$ by means of the following coupled ODEs

$$\dot{\mathbf{Q}}_{\epsilon,e}(t) \;=\; M^{(eo)}\,\mathbf{x}_o =: \mathbf{f}_e(t,\mathbf{Q}_\epsilon(t))\;,\quad \mathbf{Q}_{\epsilon,e}(0)=\boldsymbol{\epsilon}_e \tag{56}$$

$$\dot{\mathbf{Q}}_{\epsilon,o}(t) \;=\; M^{(oe)}\,\mathbb{E}\langle\mathbf{m}_e\rangle_t =: \mathbf{f}_o(t,\mathbf{Q}_\epsilon(t))\;,\quad \mathbf{Q}_{\epsilon,o}(0)=\boldsymbol{\epsilon}_o\,, \tag{57}$$

where $\mathbf{f}(t,\mathbf{Q})$ is the velocity field of the ODE. The perturbation is here introduced as an initial condition in order to have the interpolating functions in the form (44). Notice that $\mathbf{f}_e$ is constant, while $\mathbf{f}_o$ is a positive Lipschitz function of $\mathbf{Q}_\epsilon(t)\in(0,\infty)^K$ thanks to identity (28) (where N is fixed). Therefore, by Cauchy-Lipschitz's theorem, the system of ODEs (56)–(57) has a unique global solution $\mathbf{Q}_\epsilon(t)\,,\,t\in[0,1]$, whose components are positive.

By (56)–(57) we have $\Delta^{(eo)}\mathbf{q}_{\epsilon,o}(t)=\Delta^{(eo)}\mathbf{x}_o$ and $\Delta^{(oe)}\mathbf{q}_{\epsilon,e}(t)=\Delta^{(oe)}\mathbb{E}\langle\mathbf{m}_e\rangle_t$, hence:

$$\begin{aligned}&\int_0^1 dt\,\Big(\mathbf{1}_o-\mathbf{q}_{\epsilon,o}(t)\,,\;\Delta^{(oe)}(\mathbf{1}_e-\mathbf{q}_{\epsilon,e}(t))\Big)\\&\quad=\Big(\mathbf{1}_o-\mathbf{x}_o\,,\;\Delta^{(oe)}\Big(\mathbf{1}_e-\int_0^1 dt\,\mathbb{E}\langle\mathbf{m}_e\rangle_t\Big)\Big)\end{aligned} \tag{58}$$

and reasoning in a similar way for the other t-integrations appearing in the sum rule (43) we obtain:

$$\begin{aligned}\bar{p}_N\;&=\;\mathcal{O}(s_N)\;+\;p_{var}\Big(\mathbf{x}_o\,,\int_0^1 dt\,\mathbb{E}\langle\mathbf{m}_e\rangle_t\Big)\;+\;\int_0^1 dt\;R_\epsilon(t)\\&\geq\;\mathcal{O}(s_N)\;+\;\inf_{\mathbf{x}_e}p_{var}\,(\mathbf{x}_o,\mathbf{x}_e)\;+\;\int_0^1 dt\;R_\epsilon(t)\;,\end{aligned} \tag{59}$$

where the reminder is

$$R_\epsilon(t)\;=\;\frac{1}{2}\,\mathbb{E}\Big|\Big((\mathbf{m}_o-\mathbf{x}_o)\,,\;\Delta^{(oe)}\,(\mathbf{m}_e-\mathbb{E}\langle\mathbf{m}_e\rangle_t)\Big)\Big|_t\;. \tag{60}$$

Using Cauchy–Schwartz's inequality,

$$|R_\epsilon(t)|\;\leq\;\frac{1}{2}\,\Big\|\mu^{(oe)}\Big\|\;\mathbb{E}^{1/2}\langle|\hat{\alpha}^{(oo)}(\mathbf{m}_o-\mathbf{x}_o)|^2\rangle_t\;\mathbb{E}^{1/2}\langle|\hat{\alpha}^{(ee)}(\mathbf{m}_e-\mathbb{E}\langle\mathbf{m}_e\rangle_t)|^2\rangle_t\;, \tag{61}$$

thus, provided that the map $\boldsymbol{\epsilon}\mapsto\mathbf{Q}_\epsilon(t)$ is regular, the remainder $R_\epsilon(t)$ vanishes in $\boldsymbol{\epsilon}$-average as $N\to\infty$ by Lemma 4. To show that $\mathbf{Q}_\epsilon$ is regular we introduce the following matrix fields:

$$\Phi_\epsilon(t):=\frac{\partial\,\mathbf{Q}_\epsilon(t)}{\partial\boldsymbol{\epsilon}}\;,\quad A_\epsilon(t):=\frac{\partial\,\mathbf{f}(t,\mathbf{Q}_\epsilon(t))}{\partial\,\mathbf{Q}_\epsilon(t)} \tag{62}$$

Applying the chain rule we have:

$$\dot{\Phi}_\epsilon(t)=\frac{\partial\,\dot{\mathbf{Q}}_\epsilon(t)}{\partial\boldsymbol{\epsilon}}\;=\;A_\epsilon(t)\;\Phi_\epsilon(t)\;,\quad\Phi_\epsilon(0)=\mathbb{1}\;, \tag{63}$$

hence, by Liouville's formula (48) the Jacobian $\det(\Phi_\epsilon(t))$ is

$$\det\Big(\frac{\partial\mathbf{Q}_\epsilon}{\partial\boldsymbol{\epsilon}}(t)\Big)\;=\;\exp\Big\{\int_0^t ds\;\mathrm{Tr}\big(A_\epsilon(s)\big)\Big\}\;. \tag{64}$$

Now, using equations (56)–(57) one can compute:

$$\begin{aligned} \mathrm{Tr}\big(A_\epsilon(t)\big) \;&=\; \sum_{r=1}^{K} \big(A_\epsilon(t)\big)_{r,r} \;=\; \sum_{r\ \mathrm{odd}} \frac{\partial \big(M^{(oe)}\,\mathbb{E}\langle \mathbf{m}_e\rangle_t\big)_r}{\partial\, Q_{\epsilon,r}(t)} \\ &=\; \sum_{r\ \mathrm{odd}} \sum_{r'\mathrm{even}} M_{rr'}\, \frac{\partial\, \mathbb{E}\langle m_{r'}\rangle_t}{\partial\, Q_{\epsilon,r}(t)} \;\geq 0 \end{aligned} \tag{65}$$

where non-negativity is a consequence of the correlation inequality (28), since $Q_{\epsilon,r}(t)$ can be seen as the variance of an external field on the Nishimori line in the interpolating Hamiltonian (39). Combining (64) and (65), it follows that $\mathbf{Q}_\epsilon$ is regular, as desired.

Now, averaging on $\boldsymbol{\epsilon}$ and tanking the $\liminf_{N\to\infty}$ in inequality (59) we have

$$\liminf_{N\to\infty} \bar{p}_N \;\geq\; \inf_{\mathbf{x}_e} p_{var}\,(\mathbf{x}_o, \mathbf{x}_e) + \liminf_{N\to\infty} \mathbb{E}_{\boldsymbol{\epsilon}} \int_0^1 dt\; R_\epsilon(t)\,. \tag{66}$$

The last term vanishes by Fubini's theorem, dominated convergence and Lemma 4. Finally, optimizing w.r.t. $\mathbf{x}_o$ we get:

$$\liminf_{N\to\infty} \bar{p}_N \geq \sup_{\mathbf{x}_o} \inf_{\mathbf{x}_e} p_{var}\,(\mathbf{x}_o, \mathbf{x}_e)\;. \tag{67}$$

Upper bound. Now, we set

$$\dot{\mathbf{Q}}_{\epsilon,e}(t) = M^{(eo)} F(M^{(oe)}\mathbb{E}\langle \mathbf{m}_e\rangle_t + \mathbf{h}_o)\,, \quad \mathbf{Q}_{\epsilon,e}(0) = \boldsymbol{\epsilon}_e \tag{68}$$

$$\dot{\mathbf{Q}}_{\epsilon,o}(t) = M^{(oe)}\mathbb{E}\langle \mathbf{m}_e\rangle_t\,, \quad \mathbf{Q}_{\epsilon,o}(0) = \boldsymbol{\epsilon}_o\,. \tag{69}$$

In (68) the application of F, defined in (37), to the vector $M^{(oe)}\mathbb{E}\langle \mathbf{m}_e\rangle_t + \mathbf{h}_o$ has to be understood as component-wise. For future convenience let us set

$$\mathcal{D}(\mathbf{x}, \mathbf{h}) := \mathrm{diag}\left\{F'\left((M\mathbf{x})_r + h_r\right)\right\}_{r=1,\dots,K}\,. \tag{70}$$

With a slight abuse of notation we will stress the dependence of $\mathcal{D}^{(oo)}(\mathbf{x}, \mathbf{h})$ and $\mathcal{D}^{(ee)}(\mathbf{x}, \mathbf{h})$ on the even and odd components of $\mathbf{x}$ respectively as follows

$$\mathcal{D}^{(oo)}(\mathbf{x}, \mathbf{h}) \equiv \mathcal{D}^{(oo)}(\mathbf{x}_e, \mathbf{h})\,, \quad \mathcal{D}^{(ee)}(\mathbf{x}, \mathbf{h}) \equiv \mathcal{D}^{(ee)}(\mathbf{x}_o, \mathbf{h})\,. \tag{71}$$

$M^{(eo)} F(M^{(oe)}\mathbb{E}\langle \mathbf{m}_e\rangle_t + \mathbf{h}_o)$ is a positive function of $\mathbf{Q}_\epsilon(t)$ with bounded derivatives for fixed N thanks to Lemma 2, indeed

$$\frac{\partial}{\partial Q_{\epsilon,r}} F(M^{(oe)}\mathbb{E}\langle \mathbf{m}_e\rangle_t + \mathbf{h}_o) = \mathcal{D}(\mathbb{E}\langle \mathbf{m}_e\rangle_t, \mathbf{h})^{(oo)} M^{(oe)} \frac{\partial \mathbb{E}\langle \mathbf{m}_e\rangle_t}{\partial Q_{\epsilon,r}}\,, \tag{72}$$

This ensures the existence of a unique global solution over $[0, 1]$ to the system of ODEs (68)–(69). Moreover, the latter implies also that the map $\boldsymbol{\epsilon} \longmapsto \mathbf{Q}_\epsilon(\cdot)$ is still regular, because F' is positive as proved in Lemma 2 and $\frac{\partial \mathbb{E}\langle \mathbf{m}_e\rangle_t}{\partial Q_{\epsilon,r}} \geq 0$ thanks again to (28). This guarantees the positivity of the trace in (64) and forces the vanishing of the remainder R_ϵ in $\boldsymbol{\epsilon}$-average by Lemma 4. Using Jensen's inequality, by the convexity of ψ we have

$$\sum_{r=1}^{K} \alpha_r \psi\left(\left(M\int_0^1 \mathbf{q}_\epsilon(t)\,dt + \mathbf{h}\right)_r\right) \leq \sum_{r=1}^{K} \alpha_r \int_0^1 \psi\left((M\mathbf{q}_\epsilon(t) + \mathbf{h})_r\right) dt \tag{73}$$

and inserting it into the sum rule (43) we get

$$\begin{aligned}\bar{p}_N &\le \mathcal{O}(s_N) + \int_0^1 dt\; p_{var}(\mathbf{F}_{\epsilon,o}(t), \mathbb{E}\langle \mathbf{m}_e\rangle_t) + \int_0^1 R_\epsilon(t)\, dt \\ &= \mathcal{O}(s_N) + \int_0^1 dt\; \inf_{\mathbf{x}_e} p_{var}(\mathbf{F}_{\epsilon,o}(t), \mathbf{x}_e) + \int_0^1 R_\epsilon(t)\, dt\,, \end{aligned} \tag{74}$$

where $\mathbf{F}_{\epsilon,o}(t) := F(M^{(oe)}\mathbb{E}\langle \mathbf{m}_e\rangle_t + \mathbf{h}_o)$ for brevity. As far as the last equality is concerned, we used the following:

$$\inf_{\mathbf{x}_e} p_{var}(\mathbf{F}_{\epsilon,o}(t), \mathbf{x}_e) = p_{var}(\mathbf{F}_{\epsilon,o}(t), \mathbb{E}\langle \mathbf{m}_e\rangle_t)\,. \tag{75}$$

This is a consequence of the convexity of p_{var} in $\mathbf{x}_e$ (see Remark 4). In fact, a computation of the gradient of p_{var} w.r.t. $\mathbf{x}_e$ evaluated at $\mathbb{E}\langle \mathbf{m}_e\rangle_t$ yields:

$$\begin{aligned}&\frac{\partial p_{var}}{\partial \mathbf{x}_e}(\mathbf{F}_{\epsilon,o}(t), \mathbf{x}_e)\bigg|_{\mathbb{E}\langle \mathbf{m}_e\rangle_t} = \frac{\Delta^{(eo)}}{2}[\mathbf{1}_o + \mathbf{F}_{\epsilon,o}(t)] \\ &\quad + \frac{\Delta^{(eo)}}{2}[-\mathbf{1}_o + \mathbf{F}_{\epsilon,o}(t)] - \Delta^{(eo)}\mathbf{F}_{\epsilon,o}(t) = 0\,, \end{aligned} \tag{76}$$

where we explicitly notice that the first term comes from the derivative of ψ (32). Then, taking the sup of p_{var} over the odd components and the $\boldsymbol{\epsilon}$-average we get:

$$\bar{p}_N \le \mathcal{O}(s_N) + \sup_{\mathbf{x}_o}\inf_{\mathbf{x}_e} p_{var}\,(\mathbf{x}_o, \mathbf{x}_e) + \mathbb{E}_{\boldsymbol{\epsilon}} \int_0^1 R_\epsilon(t)\, dt\,. \tag{77}$$

Applying Lemma 4, Fubini's theorem and dominated convergence the two bounds match after sending $N \to \infty$.

Proof of (16). Equations (27) and (28) imply that the quenched pressure is convex in each h_r. Hence it is possible to exchange the derivative w.r.t. h_r in (27) with the $N \to \infty$ limit where $\bar{\mathbf{x}}$ is differentiable in h_r (see Lemma IV.6.3 in [17]). Since for invertible Δ the optimal order parameter must be a critical point of p_{var} (see Proposition 4 below) by (32) and (14) we have that:

$$\begin{aligned}&\lim_{N\to\infty}\frac{\partial \bar{p}_N}{\partial h_r} = \frac{\partial}{\partial h_r} p_{var}(\bar{\mathbf{x}}(M,\mathbf{h}); \mu, \mathbf{h}) = \frac{\partial p_{var}}{\partial \mathbf{x}}\bigg|_{\bar{\mathbf{x}}(M,\mathbf{h})}\frac{\partial \bar{\mathbf{x}}(M,\mathbf{h})}{\partial h_r} + \frac{\partial p_{var}}{\partial h_r} = \frac{\partial p_{var}}{\partial h_r} \\ &= \alpha_r\,\psi'((M\bar{\mathbf{x}}(M,\mathbf{h}))_r + h_r) = \frac{\alpha_r}{2}\left[1 + \mathbb{E}_z \tanh\left(z\sqrt{(M\bar{\mathbf{x}})_r + h_r} + (M\bar{\mathbf{x}})_r + h_r\right)\right] \\ &= \frac{\alpha_r}{2}\,[1 + \bar{x}_r]\,. \end{aligned} \tag{78}$$

A comparison with (27) and the Nishimori identity (22) lead to the identification:

$$\lim_{N\to\infty}\mathbb{E}\langle q_r\rangle_N = \lim_{N\to\infty}\mathbb{E}\langle m_r\rangle_N = \bar{x}_r\,. \tag{79}$$

Remark 5. Assume for now that K is even. Observe that the entire proof could have been carried out also by computing all the $\inf_{\mathbf{x}_e}$ over the convex set:

$$A := \{\mathbf{x}_e \mid M^{(oe)}\mathbf{x}_e + \mathbf{h}_o \ge 0 \text{ component-wise}\} \supseteq [0,1)^{K/2}\,, \tag{80}$$

on which all the functions involved are still real and well defined. This freedom is essentially due to the convexity of p_{var} in $\mathbf{x}_e$. Indeed, p_{var} has always a critical point in the domain A for any fixed $\mathbf{x}_o \in [0,1)^{K/2}$, that must coincide with its minimum point by convexity as can be seen by direct inspection

$$\left.\frac{\partial p_{var}}{\partial \mathbf{x}_e}\right|_{\bar{\mathbf{x}}_e} = \frac{\Delta^{(eo)}}{2}\left[-\mathbf{x}_o + F(M^{(oe)}\bar{\mathbf{x}}_e + \mathbf{h}_o)\right] = 0 \;\Leftrightarrow\; \bar{\mathbf{x}}_e = [M^{(oe)}]^{-1}(F^{-1}(\mathbf{x}_o) - \mathbf{h}_o)\,. \tag{81}$$

The inequality (59), that leads to the lower bound, clearly holds also for $\mathbf{x}_e \in A \supseteq [0,1)^{K/2}$. The validity of (75) is less trivial and is due to the special choice $\mathbf{F}_{\epsilon,o}(t)$. In this case in fact, the critical point falls inside $[0,1)^{K/2}$ and this lets us extend the domain of $\mathbf{x}_e$ to A without any loss of generality thanks to the mentioned convexity in $\mathbf{x}_e$. We will see later that even with this extension the point that realizes the sup inf lies inside the cube $[0,1)^K$.

4.2. Proof of Theorem 2. For this proof we rely on Remark 5, this will ease our computations. Let us write the gradient of (14)

$$\begin{aligned}\frac{\partial p_{var}(\mathbf{x};\mu,\mathbf{h})}{\partial x_r} &= \left(\frac{\Delta}{2}(-\mathbf{x} + F(M\mathbf{x}+\mathbf{h}))\right)_r \\ &= \frac{\Delta_{r,r+1}}{2}\left[-x_{r+1} + F((M\mathbf{x})_{r+1} + h_{r+1})\right] + \frac{\Delta_{r,r-1}}{2}\left[-x_{r-1} + F((M\mathbf{x})_{r-1} + h_{r-1})\right]\,.\end{aligned} \tag{82}$$

where we have used (31). In absence of external magnetic field ($\mathbf{h} = 0$) $\mathbf{x} = 0$ is a critical point for p_{var}, namely a solution to the consistency equation obtained by equating (82) to 0.

First of all, by Remark 4 and Remark 5 we infer that the optimization w.r.t. the even components $\mathbf{x}_e$ is always stable, in the sense that there is always one optimizer once the odd components $\mathbf{x}_o$ are fixed and it belongs to A. Define now the auxiliary function:

$$\pi(\mathbf{x}_o;\mu,\mathbf{h}) := \inf_{\mathbf{x}_e \in A} p_{var}(\mathbf{x}_o,\mathbf{x}_e;\mu,\mathbf{h}) = p_{var}(\mathbf{x}_o,\bar{\mathbf{x}}_e;\mu,\mathbf{h})\,, \tag{83}$$

with $\bar{\mathbf{x}}_e$ defined in (81). The following proposition investigates the possibility to have boundary solutions to the variational problem.

Proposition 4. *Let K be even. The points $\mathbf{x}_o$ at which the $\sup_{\mathbf{x}_o} \pi(\mathbf{x}_o;\mu,\mathbf{h})$ is attained fulfill the consistency equation:*

$$\bar{\mathbf{x}}_e = F(M^{(eo)}\mathbf{x}_o + \mathbf{h}_e)\,. \tag{84}$$

As a consequence the necessary condition for $\mathbf{x}$ to realize the $\sup_{\mathbf{x}_o}\inf_{\mathbf{x}_e} p_{var}(\mathbf{x}_o,\mathbf{x}_e;\mu,\mathbf{h})$ is to be a critical point, namely to satisfy (84).

Proof. Using (81), the gradient of π is:

$$\frac{\partial \pi(\mathbf{x}_o;\mu,\mathbf{h})}{\partial \mathbf{x}_o} = \frac{\partial p_{var}}{\partial \mathbf{x}_o} + \left.\frac{\partial p_{var}}{\partial \mathbf{x}_e}\right|_{\bar{\mathbf{x}}_e}\frac{\partial \bar{\mathbf{x}}_e}{\partial \mathbf{x}_o} = \frac{\Delta^{(oe)}}{2}\left[-\bar{\mathbf{x}}_e + F(M^{(eo)}\mathbf{x}_o + \mathbf{h}_e)\right]$$

$$= \frac{\hat{\alpha}^{(oo)}}{2}\left[-F^{-1}(\mathbf{x}_o) + \mathbf{h}_o + M^{(oe)}F(M^{(eo)}\mathbf{x}_o + \mathbf{h}_e)\right]. \tag{85}$$

We start by considering the case $\mathbf{h} = 0$. One can immediately rule out the possibility that the sup is attained at the right border, i.e. $x_{2l-1} \to 1^-$ for some l, because thanks to (38) $\partial_{x_{2l-1}}\pi \to -\infty$. Then, the necessary condition for a point $\mathbf{x}_o \in [0,1)^{K/2}$ to realize the sup is that:

$$-F^{-1}(\mathbf{x}_o) + M^{(oe)}F(M^{(eo)}\mathbf{x}_o) \le 0\,, \tag{86}$$

component-wise, where equality holds for those components for which $x_{2l-1} > 0$. If we set $M_{0,1} = M_{K,K+1} = 0$, the generic $2l-1$ component of the previous is given by

$$\begin{aligned} &- F^{-1}(x_{2l-1}) + M_{2l-1,2l-2}F(M_{2l-2,2l-3}x_{2l-3} + M_{2l-2,2l-1}x_{2l-1}) \\ &+ M_{2l-1,2l}F(M_{2l,2l-1}x_{2l-1} + M_{2l,2l+1}x_{2l+1}) \end{aligned} \tag{87}$$

whence we understand that if $x_{2l-1} = 0$ the only chance for the previous to be non positive is to have also $x_{2l-3} = x_{2l+1} = 0$ because F is positive and monotonic. On the contrary, if $x_{2l-1} > 0$ first the corresponding gradient component must vanish; second by looking at the $2l+1$ component for instance

$$\begin{aligned} &- F^{-1}(x_{2l+1}) + M_{2l+1,2l+2}F(M_{2l+2,2l+3}x_{2l+3} + M_{2l+2,2l+1}x_{2l+1}) \\ &+ M_{2l+1,2l}F(M_{2l,2l+1}x_{2l+1} + M_{2l,2l-1}x_{2l-1}) \end{aligned} \tag{88}$$

we see that the last term is strictly positive. Necessarily, x_{2l+1} must be strictly positive too with the corresponding gradient component that vanishes, and so on. Similar considerations hold for x_{2l-3}. Finally, iterating these arguments, we infer that the supremum is attained at a point $\mathbf{x}_o$ such that:

$$\mathbf{x}_o = 0 \quad \text{or} \quad \bar{\mathbf{x}}_e = F(M^{(eo)}\mathbf{x}_o)\,. \tag{89}$$

The first in particular implies that also $\bar{\mathbf{x}}_e = 0$. In both cases we can say that (84) is satisfied.

When any h_r is strictly positive it is immediate to see that there is a component of (85) with a positive contribution, the corresponding component of $\mathbf{x}_o$ must then be positive. Therefore one iterates the same arguments as above obtaining again (84). In any case, by (81) the sup inf is attained at critical points of p_{var}. □

Using property (33), the Jacobian matrix of $F(M\mathbf{x}+\mathbf{h})$ is

$$DF(M\mathbf{x}+\mathbf{h}) = \mathcal{D}(\mathbf{x},\mathbf{h})M\,. \tag{90}$$

$\mathcal{D}(\mathbf{x},\mathbf{h})$, defined in (70), is diagonal, positive definite, invertible and its spectral radius is bounded by 1. From (85), an application of the Inverse Function Theorem leads to the Hessian matrix

$$\begin{aligned} \mathscr{H}_{\mathbf{x}_o}\pi &= \frac{\Delta^{(oe)}}{2}\left[-\frac{\partial\bar{\mathbf{x}}_e}{\partial\mathbf{x}_o} + \frac{\partial}{\partial\mathbf{x}_o}F(M^{(eo)}\mathbf{x}_o + \mathbf{h_e})\right] \\ &= \frac{\Delta^{(oe)}}{2}\left[-[M^{(oe)}]^{-1}[\mathcal{D}^{(oo)}(\bar{\mathbf{x}}_e,\mathbf{h})]^{-1} + \mathcal{D}^{(ee)}(\mathbf{x}_o,\mathbf{h})M^{(eo)}\right]. \end{aligned} \tag{91}$$

Thanks to the peculiar tridiagonal form of M we also have that

$$[\mathcal{D}(\mathbf{x},\mathbf{h})M]^{(oe)}[\mathcal{D}(\mathbf{x},\mathbf{h})M]^{(eo)} = [(\mathcal{D}(\mathbf{x},\mathbf{h})M)^2]^{(oo)}\,, \tag{92}$$

from which by a simple rearrangement we can write the Hessian in its final form

$$\begin{aligned}\mathscr{H}_{\mathbf{x}_o}\pi &= \frac{\hat{\alpha}^{(oo)}[\mathcal{D}(\bar{\mathbf{x}}_e,\mathbf{h})^{(oo)}]^{-1}}{2}\left[-\gg 1 + (\mathcal{D}(\mathbf{x},\mathbf{h})M)^2\right]^{(oo)} \\ &= \frac{[\alpha^{(oo)}]^{1/2}[\mathcal{D}^{(oo)}]^{-1/2}}{2}\left[-\gg 1 + S^{(oo)}\right][\alpha^{(oo)}]^{1/2}[\mathcal{D}^{(oo)}]^{-1/2}\end{aligned} \tag{93}$$

with

$$S^{(oo)} := [\mathcal{D}^{(oo)}]^{1/2}[\hat{\alpha}^{(oo)}]^{-1/2}\Delta^{(oe)}\mathcal{D}^{(ee)}[\hat{\alpha}^{(ee)}]^{-1}\Delta^{(eo)}[\hat{\alpha}^{(oo)}]^{-1/2}[\mathcal{D}^{(oo)}]^{1/2} \tag{94}$$

where for brevity we have neglected all the dependencies after the second equality in (93) and used (92). (93) uses only symmetric matrices in order to make manifest the global sign of the Hessian. It remains to show that the spectral radius of $S^{(oo)}$ is controlled by that of $[M^2]^{(oo)}$. $S^{(oo)}$ is symmetric because $\Delta^{(oe)} = [\Delta^{(eo)}]^T$, thus its spectral radius coincides with the matrix norm induced by the Euclidean scalar product. Then by norms sub-multiplicativity and matrix similarity one easily gets

$$\begin{aligned}\rho\left(S^{(oo)}\right) &\leq \rho\left(\mathcal{D}^{(oo)}\right)\rho\left([\hat{\alpha}^{(oo)}]^{-1/2}\Delta^{(oe)}\mathcal{D}^{(ee)}[\hat{\alpha}^{(ee)}]^{-1}\Delta^{(eo)}[\hat{\alpha}^{(oo)}]^{-1/2}\right) \\ &\leq \rho\left(M^{(oe)}\mathcal{D}^{(ee)}M^{(eo)}\right) = \rho\left(\mathcal{D}^{(ee)}M^{(eo)}M^{(oe)}\right) \\ &= \rho\left([\mathcal{D}^{(ee)}]^{1/2}[\hat{\alpha}^{(ee)}]^{-1/2}\Delta^{(eo)}\hat{\alpha}^{(oo)-1}\Delta^{(oe)}[\hat{\alpha}^{(ee)}]^{-1/2}[\mathcal{D}^{(ee)}]^{1/2}\right)\,.\end{aligned} \tag{95}$$

Iterating the same arguments we get to

$$\rho\left(S^{(oo)}\right) \leq \rho\left(M^{(eo)}M^{(oe)}\right) = \rho\left(M^{(oe)}M^{(eo)}\right) < 1\,, \tag{96}$$

where the last equality follows again by matrix similarity. The previous implies that the matrix $[-\gg 1 + S]^{(oo)}$ in (93) is negative definite, making $\mathscr{H}_{\mathbf{x}_o}\pi$ negative definite too, and hence π is concave under the hypothesis $\rho([M^2]^{(oo)}) < 1$. In turn, this ensures uniqueness of the solution to the consistency equation (84) and to the variational problem (13). In particular when $\mathbf{h} = 0$, $\mathbf{x} = 0$ is the unique solution.

Conversely, for $\mathbf{h} = 0$ and $\rho([M^2]^{(oo)}) > 1$ the Hessian has at least one positive eigenvalue at the origin $\mathbf{x}_o = 0$, but this is in general not enough to ensure $\mathbf{x}_o = 0$ does not realize the sup anymore. One has to check that there is a direction of increment of π that intersects the cube $[0,1)^{K/2}$, otherwise the system could remain stuck on the border at $\mathbf{x}_o = 0$ due to the positivity constraints on the variables.

It is easy to see that $[M^2]^{(oo)}$ is irreducible, because its associated graph is strongly connected, and it has non negative entries. Hence, by Perron–Frobenius Theorem the eigenvector $\mathbf{v}$ relative to the largest eigenvalue $\rho([M^2]^{(oo)})$ is component-wise strictly positive, thus pointing inside the cube, and by a Taylor expansion around $\mathbf{x}_o = 0$ we have:

$$\pi(\epsilon\mathbf{v};\mu,0) - \pi(0;\mu,0) = \frac{\epsilon^2}{2}\left(\mathbf{v},\frac{\hat{\alpha}^{(oo)}}{2}\mathbf{v}\right)\left[-1+\rho([M^2]^{(oo)})\right] + o(\epsilon^2) \tag{97}$$

that is positive form small enough $\epsilon > 0$. Finally, by Proposition 4 the solution shifts in favour of a point $\mathbf{x} = \bar{\mathbf{x}}(M) \in (0,1)^K$.

4.3. Proof of Proposition 1. Proposition 1 relies on an algebraic lemma, which we write here for convenience. Its proof can be found in [3] (see Lemma 1 therein).

Lemma 5. *Let* $P \geq 2$, $x_1, \dots, x_P \geq 0$ *and* $b_1, \dots, b_{P-1} \geq 0$. *Set* $S \equiv \sum_{p=1}^{P} x_p$ *and* $B \equiv \max_{p=1,\dots,P-1} b_p$. *Then:*

$$\sum_{p=1}^{P-1} b_p \, x_p \, x_{p+1} \;\leq\; \frac{B\, S^2}{4} \,. \tag{98}$$

Moreover we have equality in (98) *if and only if one of the following conditions is verified:*

(a) *there exists* $p^* \in \{1, \dots, P-1\}$ *such that*

$$x_{p^*} \;=\; x_{p^*+1} \;=\; \frac{S}{2} \quad , \quad b_{p^*} = B \;; \tag{99}$$

(b) *there exists* $p^* \in \{2, \dots, P-1\}$ *such that*

$$x_{p^*} \;=\; x_{p^*-1} + x_{p^*+1} \;=\; \frac{S}{2} \quad , \quad b_{p^*-1} = b_{p^*} = B \,. \tag{100}$$

Proof of Proposition 1. Denote by ρ the spectral radius of the matrix $[M^2]^{(oo)}$. We have

$$\rho \;\leq\; \left\| [M^2]^{(oo)} \right\|_\infty \,. \tag{101}$$

As $[M^2]^{(oo)}$ is a tridiagonal matrix, its ∞-norm can be easily computed leading to

$$\left\| [M^2]^{(oo)} \right\|_\infty \;=\; \max_r \sum_s (M^2)_{2r-1,2s-1} \;=\; \max_r \sum_{p=2r-3}^{2r} b_p^{(r)} \, \alpha_p \, \alpha_{p+1} \;\leq\; \frac{\widehat{\mu}^2}{4} \,, \tag{102}$$

where we set $\widehat{\mu}^2 \equiv \max_r \mu^2_{r,r+1}$ and for every r, p

$$\begin{aligned} b_p^{(r)} \;\equiv\; & \delta_{p,\,2r-3} \; \mu_{2r-3,\,2r-2} \; \mu_{2r-2,\,2r-1} \;+\; \delta_{p,\,2r-2} \; \mu^2_{2r-2,\,2r-1} \;+ \\ & +\; \delta_{p,\,2r-1} \; \mu^2_{2r-1,\,2r} \;+\; \delta_{p,\,2r} \; \mu_{2r-1,\,2r} \; \mu_{2r,\,2r+1} \,. \end{aligned} \tag{103}$$

For convenience we set $\alpha_p \equiv 0$ for $p \notin \{1, \dots, K\}$ and $\mu_{p,p+1} \equiv 0$ for $p \notin \{1, \dots, K-1\}$. The last inequality in (102) follows by Lemma 5, since $b_p^{(r)} \leq \widehat{\mu}^2$ and $\sum_p \alpha_p = 1$. Therefore $\rho \leq \frac{\widehat{\mu}^2}{4}$ combining inequalities (101), (102).

Now assume that $\rho = \frac{\widehat{\mu}^2}{4}$. In particular the inequality in (102) must be saturated, hence there exists r such that

$$\sum_{p=2r-3}^{2r} b_p^{(r)} \, \alpha_p \, \alpha_{p+1} \;=\; \frac{\widehat{\mu}^2}{4} \,. \tag{104}$$

Then by Lemma 5, condition (19) or (20) must be verified.

Vice-versa assume that condition (19) or (20) holds true. In this case notice that many of the α_r's are zero, since $\sum_{r=1}^{K} \alpha_r = 1$. Thus the matrix $[M^2]^{(oo)}$ notably simplifies and one can check directly that $\frac{\widehat{\mu}^2}{4}$ is (the only non-zero) eigenvalue. This proves $\rho = \frac{\widehat{\mu}^2}{4}$. □

Remark 6. It is not difficult to realize that Theorem 1 holds also when $\alpha_r \to 0$ for some r. Indeed, by (61) and Lemma 4 we see that it is sufficient to require:

$$\alpha_r^2 \, \mathbb{E}\Big\langle \Big(\mathcal{L}_r - \mathbb{E}\langle \mathcal{L}_r \rangle_{N,t}^{(\epsilon)} \Big)^2 \Big\rangle_{N,t}^{(\epsilon)} \longrightarrow 0 \quad \text{as } N \to \infty \,. \tag{105}$$

The proof of Lemma 4 consists in showing that (see inequalities (A.11) and (A.24) in [2]):

$$\begin{aligned} &\alpha_r^2 \, \mathbb{E}_{\boldsymbol{\epsilon}} \mathbb{E}\Big\langle \Big(\mathcal{L}_r - \langle \mathcal{L}_r \rangle_{N,t}^{(\epsilon)} \Big)^2 \Big\rangle_{N,t}^{(\epsilon)} = O\left(\frac{\alpha_r}{N s_N^K} \right) \\ &\alpha_r^2 \, \mathbb{E}_{\boldsymbol{\epsilon}} \mathbb{E}\left[\Big(\langle \mathcal{L}_r \rangle_{N,t}^{(\epsilon)} - \mathbb{E}\langle \mathcal{L}_r \rangle_{N,t}^{(\epsilon)} \Big)^2 \right] = O\left(\frac{1}{s_N^{4K/3} N^{1/3}} \right) . \end{aligned} \tag{106}$$

The previous equalities both vanish in the thermodynamic limit with the choice $s_N \sim N^{-1/16K}$ for instance, independently on α_r. Hence the remainder of the interpolation in the proof still goes to 0 with no variation in the hypothesis.

When a form factor, say α_r, vanishes the related component of the order parameter x_r disappears from the variational pressure (14). Moreover, if the corresponding L_r is an intermediate layer one can see that the system decouples into two independent DBMs because the effective interaction matrix Δ becomes block diagonal and the convex ψ-term related to the mentioned layer is weighed by α_r. The global variational pressure is thus constant in x_r in the thermodynamic limit.

4.4. Proof of Theorem 3. Uniqueness of the solution of the consistency equation for positive external fields can be proven adapting the strategy used in [3], where the replica symmetric equation of a deep Boltzmann machines was proved to admit a unique solution when the couplings and the external fields are centred Gaussian random variables. In particular the layers structure permits to "decouple" the interactions as shown in the following

Remark 7. The consistency equation (21) is equivalent to the following:

$$\begin{cases} x_r \;=\; \mathbb{E} \tanh\left(z \sqrt{\Theta_r(a)\, x_r + h_r} \;+\; \Theta_r(a)\, x_r + h_r \right) & r = 1, \dots, K \\ \alpha_r \, x_r \, a_r \;=\; \alpha_{r+1} \, x_{r+1} & r = 1, \dots, K-1 \end{cases} \tag{107}$$

where we have introduced the auxiliary variables $a_1, \dots, a_{K-1} > 0$ and the functions

$$\Theta_r(a) \;\equiv\; \begin{cases} \Delta_{12}\, a_1 & \text{for } r = 1 \\ \dfrac{\Delta_{r,r-1}}{a_{r-1}} \;+\; \Delta_{r,r+1}\, a_r & \text{for } r = 2, \dots, K-1 \\ \dfrac{\Delta_{K-1,K}}{a_{K-1}} & \text{for } r = K \end{cases} . \tag{108}$$

Indeed, using the definition of the matrix M, it can be easily verified that $(M\mathbf{x})_r = \Theta_r(a)\, x_r$ for $r = 1, \dots, K$, for a satisfying the second relation in (107).

The proof of Theorem 3 relies on the following

Lemma 6. *Let z be a standard Gaussian random variable. For every $t, h > 0$ the equation*

$$x = \mathbb{E} \tanh\left(z\sqrt{t\,x+h} + t\,x+h\right) \tag{109}$$

has a unique positive solution that we denote by $x = \bar{x}(t,h) > 0$. Moreover $\bar{x}$ is strictly increasing as a function of both $t > 0$ and $h > 0$.

Proof of Theorem 3. Equation (109) rewrites as $x = F(t\,x+h)$, where $F(h) \equiv \mathbb{E}\tanh(z\sqrt{h} + h)$. By Lemma 2, F takes values in (0,1), is strictly increasing and concave. It follows that equation (109) admits a unique solution in (0, 1) and in particular we can show that the function $f(x) \equiv \frac{1}{x} F(t\,x+h)$ is strictly decreasing for $x > 0$. Indeed by Lemma 2 we have:

$$x^2\, f'(x) = t\,x\,F'(t\,x+h) - F(t\,x+h) < 0 \text{ in } x = 0\,, \tag{110}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x^2\, f'(x)\right) = t^2\,x\,F''(t\,x+h) < 0 \tag{111}$$

hence

$$x^2\, f'(x) = t\,x\,F'(t\,x+h) - F(t\,x+h) < 0 \quad \forall\, x > 0\,. \tag{112}$$

Now denoting by $\bar{x}(t,h)$ the unique positive solution of equation (109), we can prove its monotonicity with respect to both parameters by differentiating the self-consistent equation

$$\bar{x}(t,h) = F\left(t\,\bar{x}(t,h)+h\right)\,, \tag{113}$$

which leads to

$$\left(1 - t\,F'(t\,\bar{x}+h)\right)\frac{\mathrm{d}\bar{x}}{\mathrm{d}t} = \bar{x}\,F'(t\,\bar{x}+h) \tag{114}$$

$$\left(1 - t\,F'(t\,\bar{x}+h)\right)\frac{\mathrm{d}\bar{x}}{\mathrm{d}h} = F'(t\,\bar{x}+h)\,. \tag{115}$$

Lemma 2 ensures that (114), (115) are positive quantities, hence to conclude it suffices to show that $1 - t\,F'(t\,\bar{x}+h) > 0$. Indeed, dividing the inequality (112) by x, evaluating it at $x = \bar{x}(t,h)$ and using the self-consistent equation (113), one finds precisely:

$$0 > t\,F'(t\,\bar{x}+h) - \frac{F(t\,\bar{x}+h)}{\bar{x}} = t\,F'(t\,\bar{x}+h) - 1\,. \tag{116}$$

□

Proof of Theorem 3. By Lemma 6, the first line of (107) is equivalent to:

$$x_r = \bar{x}\left(\Theta_r(a), h_r\right) \quad \forall\, r = 1, \ldots, K \tag{117}$$

where $\bar{x}$ is uniquely defined and strictly increasing with respect to both its arguments. On the other hand the second line of (107) rewrites as:

$$\alpha_1\, x_1\, a_1 \cdots a_r = \alpha_{r+1}\, x_{r+1} \quad \forall\, r = 1, \ldots, K-1\,. \tag{118}$$

It is convenient to set $X_1(a_1) \equiv \alpha_1 \, \bar{x}\left(\Theta_1(a)\,,\, h_1\right) = \alpha_1 \, \bar{x}\big(\Delta_{1,2}\, a_1\,,\, h_1\big)$ and for $r \geq 2$

$$X_r\left(\frac{1}{a_{r-1}}\,,\, a_r\right) \equiv \alpha_r \, \bar{x}\left(\Theta_r(a)\,,\, h_r\right) = \alpha_r \, \bar{x}\left(\frac{\Delta_{r,r-1}}{a_{r-1}} + \Delta_{r,r+1}\, a_r\,,\, h_r\right). \quad (119)$$

Therefore equation (107) is equivalent to the following:

$$X_1(a_1)\, a_1 \cdots a_r \;=\; X_{r+1}\left(\frac{1}{a_r}\,,\, a_{r+1}\right) \quad \forall\, r = 1, \ldots, K-1\,. \quad (120)$$

We will show by induction on $r \geq 1$ that for any given $a_{r+1} \geq 0$ there exists a unique $a_r = \bar{a}_r(a_{r+1}) > 0$ such that

$$\begin{cases} a_{r-1} \;=\; \bar{a}_{r-1}(a_r) \\ \vdots \\ a_1 \;=\; \bar{a}_1(a_2) \\ X_1(a_1)\, a_1 \,\cdots\, a_{r-1}\, a_r \;=\; X_{r+1}\left(\dfrac{1}{a_r}\,,\, a_{r+1}\right) \end{cases} \quad (121)$$

and moreover $\bar{a}_r$ is a strictly increasing function with respect to a_{r+1}. The uniqueness of solution of (120) will follow immediately by stopping the induction at $r = K-1$ and choosing $a_K = 0$ and the Theorem will be proven thanks to Remark 7.

• Case $r = 1$: given $a_2 \geq 0$, let's consider the equation

$$X_1(a_1)\, a_1 \;=\; X_2\left(\frac{1}{a_1}, a_2\right). \quad (122)$$

By Lemma 6 the left-hand side of (122) is a strictly increasing function of $a_1 > 0$ and takes all the values in the interval $(0, \infty)$, while the right-hand side is a decreasing function of $a_1 > 0$ and takes non-negative values. Therefore there exists a unique $a_1 = \bar{a}_1(a_2) > 0$ solution of (122). Now taking derivatives on both sides of (122) and using again Lemma 6, one finds:

$$\frac{d\bar{a}_1}{da_2} \;=\; \frac{\partial}{\partial a_2} X_2\Big(\frac{1}{a_1}, a_2\Big)\left[\frac{\partial}{\partial a_1}\big(X_1(a_1)\, a_1\big) - \frac{\partial}{\partial a_1} X_2\Big(\frac{1}{a_1}, a_2\Big)\right]^{-1}_{\big|a_1 = \bar{a}_1(a_2)} > 0 \quad (123)$$

hence $\bar{a}_1$ is a strictly increasing function of a_2.

• For $r > 1$, $r-1 \Rightarrow r$. Fix $a_{r+1} \geq 0$. By inductive hypothesis $\bar{a}_1, \ldots, \bar{a}_{r-1}$ are well-defined and strictly increasing functions. Defining the composition $A_l \equiv \bar{a}_l \circ \cdots \circ \bar{a}_{r-1}$ for every $l = 1, \ldots, r-1$, equation (121) rewrites as:

$$\big(X_1 \circ A_1\big)(a_r)\; A_1(a_r) \,\cdots\, A_{r-1}(a_r)\; a_r \;=\; X_{r+1}\left(\frac{1}{a_r}, a_{r+1}\right). \quad (124)$$

By inductive hypothesis and Lemma 6, the left-hand side of (124) is a strictly increasing function of $a_r > 0$ and takes all the values in the interval $(0, \infty)$, while the right hand-side of (124) is a decreasing function of $a_r > 0$ and takes non-negative values. Therefore

for every $a_{r+1} \geq 0$ there exists a unique $a_r = \bar{a}_r(a_{r+1}) > 0$ solution of (124). Now taking derivatives on both sides of (124) one finds:

$$\frac{d\bar{a}_r}{da_{r+1}} = \frac{\partial}{\partial a_{r+1}} X_{r+1}\Big(\frac{1}{a_r}, a_{r+1}\Big) \cdot \cdot \Big[\frac{\partial}{\partial a_r}\Big(\big(X_1 \circ A_1\big)(a_r)\ A_1(a_r) \cdots A_{r-1}(a_r)\ a_r\Big) - \frac{\partial}{\partial a_r} X_{r+1}\Big(\frac{1}{a_r}, a_{r+1}\Big)\Big]^{-1}_{\Big| a_r = \bar{a}_r(a_{r+1})} \tag{125}$$

which, using again the inductive hypothesis and Lemma 6, entails that $\bar{a}_r$ is a strictly increasing function of a_{r+1} . □

5. Conclusions and Perspectives

In this work we have solved the K-layer deep restricted Boltzmann machine on the Nishimori line which is an instance of a non-convex multi-species model. The solution consists in the computation of the pressure in the thermodynamic limit which is expressed in terms of an ordinary min-max variational principle over K real positive numbers. The properties of the optimizer show the presence of a phase transition related to the interaction strength and to the relative size of each layer defining the geometry of the system. In particular we discovered that the geometry of the system may tune the phase transition.

A possible way to investigate the model for general values of the parameters would be to test the stability of our results when the system is in a neighborhood of the Nishimori line. We plan to perturb the distribution (2) and check under which conditions the replica symmetry property breaks down.

After the completion of this work, paper [28] was brought to our attention where the mutual information for a wide class of inference problems is solved by means of a variational principle. While it is possible to obtain our model as an instance of the one considered there, the variational principle presented has no clear correspondence to ours. We finally mention that a subsequent work [14] contains a general result that extends the one in the present paper. In particular the authors compute the limiting free energy with a Hamilton-Jacobi approach which proves to be effective also when dealing with lack of convexity in the interactions. On the other hand, the simplicity of our setting allows us to carry out a thorough study of the variational formula by locating the phase transition and investigating its dependency on the geometry of the system as in Theorem 2, Proposition 1 and Theorem 3.

Acknowledgement. The authors thank Jean Barbier and Francesco Guerra for interesting discussions. We acknowledge Jean-Cristophe Mourrat for bringing reference [28] to our attention and an anonymous referee for poiting out to us the preprint [14]. P.C. acknowledges support from EU project 952026-Humane-AI-Net. D.A. and E.M. acknowledge support from Progetto Alma Idea 2018, Università di Bologna.

Funding Open access funding provided by Alma Mater Studiorum - Universitá di Bologna within the CRUI-CARE Agreement.

References

1. Alberici, D., Barra, A., Contucci, P., Mingione, E.: Annealing and replica-symmetry in deep Boltzmann machines. J. Stat. Phys. **180**, 665–677 (2020)
2. Alberici, D., Camilli, F., Contucci, P., Mingione, E.: The multi-species mean-field spin-glass on the Nishimori line. J. Stat. Phys. **182**, 1–20 (2020)
3. Alberici, D., Contucci, P., Mingione, E.: Deep Boltzmann Machines: rigorous results at arbitrary depth. Annales Institut Henri Poincaré (to appear) (2021)
4. Auffinger, A., Chen, W.K.: Free energy and complexity of spherical bipartite models. J. Stat. Phys. **157**, 40–59 (2014)
5. Baik, J., Lee, J.O.: Free energy of bipartite spherical Sherrington–Kirkpatrick model. Ann. Inst. Henri Poincaré **56**, 2897–2934 (2020)
6. Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T., Zdeborová, L.: Mutual information for symmetric rank-one matrix estimation: a proof of the replica formula. Adv. Neural Inf. Process. Syst. **29**, 424–432 (2016)
7. Barbier, J., Dia, M., Macris, N., Krzakala, F., Zdeborová, L.: Rank-one matrix estimation: analysis of algorithmic and information theoretic limits by the spatial coupling method. arXiv:1812.02537 (2018)
8. Barbier, J., Macris, N.: The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. Probab. Theory Relat. Fields **174**, 1133–1185 (2019)
9. Barbier, J., Macris, N., Miolane, L.: The layered structure of tensor estimation and its mutual information. In: 55th Annual Allerton Conference on Communication, Control, and Computing (2017)
10. Barbier, J., Panchenko, D.: Strong replica symmetry in high-dimensional optimal Bayesian inference. arXiv:2005.03115 (2020)
11. Barra, A., Contucci, P., Mingione, E., Tantari, D.: Multi-species mean field spin glasses. Rigorous results. Ann. Inst. Henri Poincaré **16**, 691–708 (2013)
12. Barra, A., Genovese, G., Guerra, F.: Equilibrium statistical mechanics of bipartite spin systems. J. Phys. A Math. Theor. **44**, 245002 (2011)
13. Bates, E., Sloman, L., Sohn, Y.: Replica symmetry breaking in multi-species Sherrington–Kirkpatrick model. J. Stat. Phys. **174**, 333–350 (2018)
14. Chen, H.B., Mourrat, J.C., Xia, J.: Statistical inference of finite-rank tensors. arXiv e-prints (2021)
15. Contucci, P., Giardinà, C.: Perspectives on Spin Glasses. Cambridge University Press, Cambridge (2012). https://doi.org/10.1017/CBO9781139049306
16. Contucci, P., Morita, S., Nishimori, H.: Surface terms on the Nishimori line of the Gaussian Edwards-Anderson model. J. Stat. Phys. **122**, 303–312 (2005)
17. Ellis, R.: Entropy, Large Deviations, and Statistical Mechanics. Springer, Berlin (2006)
18. Genovese, G.: Minimax formula for the replica symmetric free energy of deep restricted Boltzmann machines. arXiv:2005.09424 (2020)
19. Guerra, F.: Broken replica symmetry bounds in the mean field spin glass model. Commun. Math. Phys. **233**, 1–12 (2003)
20. Guerra, F., Toninelli, F.L.: The thermodynamic limit in mean field spin glass models. Commun. Math. Phys. **230**, 71–79 (2002)
21. Morita, S., Nishimori, H., Contucci, P.: Griffiths inequalities for the Gaussian spin glass. J. Phys. A Math. Gen. **37**, L203 (2004)
22. Morita, S., Nishimori, H., Contucci, P.: Griffiths inequalities in the Nishimori line. Prog. Theor. Phys. Suppl. **157**, 73–76 (2005)
23. Mourrat, J.C.: Free energy upper bound for mean-field vector spin glasses. arXiv:2010.0911 (2020)
24. Mourrat, J.C.: Nonconvex interactions in mean-field spin glasses. arXiv:2004.01679 (2020)
25. Nishimori, H.: Internal energy, specific heat and correlation function of the bond-random Ising model. Prog. Theor. Phys. **66**, 1169–1181 (1981)
26. Nishimori, H.: Statistical Physics of Spin Glasses and Information Processing: an Introduction. Oxford University Press, Oxford (2001)
27. Panchenko, D.: The free energy in a multi-species Sherrington–Kirkpatrick model. Ann. Probab. **43**, 3494–3513 (2015)

28. Reeves, G.: Information-theoretic limits for the matrix tensor product. IEEE J. Sel. Areas Inf. Theory **1**, 777–798 (2020)
29. Talagrand, M.: Mean Field Models for Spin Glasses: Volume I: Basic Examples. Springer, Berlin (2010)

Communicated by S. Chatterjee