Physica A 633 (2024) 129372

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

About the de Almeida–Thouless line in neural networks

L. Albanese [a,b,d,e,*], A. Alessandrelli [c,e], A. Annibale [b], A. Barra [a,e]

[a] Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento, Lecce, Italy
[b] Department of Mathematics, King's College London, Strand, London WC2R 2LS, UK
[c] Dipartimento di Informatica, Università di Pisa, Lungarno Antonio Pacinotti, 43, 56126, Pisa, Italy
[d] Scuola Superiore ISUFI, Università del Salento, Lecce, Italy
[e] Istituto Nazionale di Fisica Nucleare, Sezione di Lecce, Italy

ARTICLE INFO

Keywords:
Neural networks
AT line
de Almeida–Thouless line
RSB
Hopfield model
Dense associative memories

ABSTRACT

In this work we present a rigorous and straightforward method to detect the onset of the instability of replica-symmetric theories in information processing systems, which does not require a full replica analysis as in the method originally proposed by de Almeida and Thouless for spin glasses. The method is based on an expansion of the free-energy obtained within one-step of replica symmetry breaking (RSB) around the RS value. As such, it requires solely continuity and differentiability of the free-energy and it is robust to be applied broadly to systems with quenched disorder. We apply the method to the Hopfield model and to neural networks with multi-node Hebbian interactions, as case studies. In the appendices we test the method on the Sherrington–Kirkpatrick and the Ising P-spin models, recovering the AT lines known in the literature for these models, as a special limit, which corresponds to assuming that the transition from the RS to the RSB phase can be obtained by varying continuously the order parameters. Our method provides a generalization of the AT approach, which does not rely on this limit and can be applied to systems with discontinuous phase transitions, as we show explicitly for the spherical P-spin model, recovering the known RS instability line.

1. Introduction

Replica symmetry breaking in neural networks has attracted increasing attention in recent years [1–5], however there is, as yet, no general broken replica-symmetry theory for these systems and no simple method to systematically detect their transition from the replica-symmetric (RS) to the replica-symmetry-broken (RSB) phase.

While the first point is still out of reach, the second question can be addressed by adapting approaches originally developed for spin glasses. Indeed, the instability line of the RS phase in the Sherrington–Kirkpatrick (SK) spin glass model was derived by de Almeida and Thouless (AT) many decades ago [6], using a method based on replicas. Since their seminal work, rigorous techniques have been developed and tested in archetypical mean-field as well as short-ranged spin glass models, by many researchers (see e.g. [7–14]).

As neural networks are particular realizations of spin glasses, it is quite natural to ask if we can devise a systematic method to derive the RS instability line also for these systems. In this work we answer affirmatively to this question, using the Hopfield model of neural networks and a model of dense associative memory, which extends Hebbian learning to multi-node interactions, as case studies. To this purpose, we devise a method inspired by the approach proposed by Toninelli in [15], which builds on Guerra's work on broken replica-symmetry bounds [16]. As a technical note we remark that at difference with conventional spin glasses, here we

* Corresponding author at: Dipartimento di Matematica e Fisica "Ennio De Giorgi", Università del Salento, Lecce, Italy.
E-mail address: linda.albanese@unisalento.it (L. Albanese).

https://doi.org/10.1016/j.physa.2023.129372
Received 11 March 2023; Received in revised form 12 November 2023

focus on the RS instability in the parameter space (α, T) where α is the storage load of the network and T is the noise level, rather than in the space (h, T) (i.e. magnetic field, temperature) conventionally used in spin glasses.

For the Hopfield model, our method recovers the instability line obtained by Coolen [17] using the AT approach, as a special limit, which corresponds to assuming a continuous transition from the RS to the RSB phase, in the order parameters. Therefore, our method provides a generalization of the AT approach, which can be applied to systems with a discontinuous transition from the RS to the 1RSB phase. Another advantage of our method, when compared to the involved calculations of the AT method in the Hopfield model [17], is its remarkable simplicity. This allows for straightforward application to more complex neural network models, such as dense associative memories with P-node interactions. We supplement the results in the main text with Appendix A where we show our method at work on conventional spin-glass models, namely the Sherrington–Kirkpatrick model, the Ising P-spin and the spherical P-spin model, the latter providing an example of system which exhibits a discontinuous phase transition from the RS to the RSB phase. In all cases, we retrieve the AT lines known for these models [6,17–19] in a specific limit, confirming the validity of our approach as a generalization of the AT method. As expected from the decomposition theorem of multi-node Hebbian networks proved in [1], for dense associative memories with P-node interactions we retrieve the instability line of the Ising P-spin model derived in Appendix A. Appendices B and C provide further technical details.

2. The Hopfield model

In this section we illustrate the method for the Hopfield model with N Ising neurons $\sigma_i \in \{1, -1\}$, $i = 1, \dots, N$ and $K = \alpha N$ stored patterns $\boldsymbol{\xi}^\mu$, $\mu = 1, \dots, K$. Each pattern $\boldsymbol{\xi}^\mu$ is a sequence of N Rademacher entries (i.e. Bernoulli variables) ξ_i^μ, $i = 1, \dots, N$, with distribution

$$\mathbb{P}(\xi_i^\mu) = \frac{1}{2}\left(\delta_{\xi_i^\mu,+1} + \delta_{\xi_i^\mu,-1}\right). \tag{2.1}$$

The Hamiltonian of the model is

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{i,j=1,1}^{N,N}\sum_{\mu=1}^{K}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j \tag{2.2}$$

and we denote the associated Boltzmann factor, at inverse temperature $\beta = 1/T$, as

$$\mathcal{B}_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = \frac{e^{-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi})}}{Z}, \qquad Z = \sum_{\boldsymbol{\sigma}} e^{-\beta H_N(\boldsymbol{\sigma}|\boldsymbol{\xi})}. \tag{2.3}$$

In the so-called 'retrieval' phase, the equilibrium local configurations are correlated only with a single pattern, say ν. As the couplings $J_{ij} = \sum_{\mu=1}^{K}\xi_i^\mu\xi_j^\mu$ are symmetric w.r.t. permutations of the patterns, it is assumed without loss of generality that $\nu = 1$. It is then convenient to define as the order parameters of the system, the so-called *Mattis magnetization*

$$m(\boldsymbol{\sigma}) := \frac{1}{N}\sum_{i=1}^{N}\xi_i^1\sigma_i \tag{2.4}$$

which quantifies the alignment of the system configuration $\boldsymbol{\sigma}$ with the retrieval pattern $\boldsymbol{\xi}^1$, and the *two-replica overlap*

$$q(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}) := \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)} \tag{2.5}$$

which quantifies the correlations between two configurations $\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}$ of the system, with the same realization of the patterns (i.e. quenched disorder).

The RS analysis assumes that the order parameters m and q self-average around their equilibrium values $\bar{m}$ and $\bar{q}$, in the thermodynamic limit, namely

$$\lim_{N\to+\infty} P_N(m) = \delta(m - \bar{m}), \tag{2.6}$$

$$\lim_{N\to+\infty} P'_N(q) = \delta(q - \bar{q}), \tag{2.7}$$

where $P_N(m) = \mathbb{E}_\xi \sum_{\boldsymbol{\sigma}} \mathcal{B}_N(\boldsymbol{\sigma}|\boldsymbol{\xi})\delta(m - m(\boldsymbol{\sigma}))$ and $P'_N(q) = \mathbb{E}_\xi \sum_{\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}} \mathcal{B}_N(\sigma^{(1)}|\boldsymbol{\xi})\mathcal{B}_N(\boldsymbol{\sigma}^{(2)}|\boldsymbol{\xi})\delta(q - q_{12}(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}))$, with $\mathbb{E}_\xi$ denoting the expectation over the pattern distribution (or 'quenched' disorder). Under this assumption, the free-energy, averaged over the pattern distribution, f, is given by (see [20])

$$\begin{aligned}-\beta f_{RS}(\bar{m}, \bar{q}|\beta, \alpha) = \ln 2 - \frac{\alpha}{2}\ln(1 - \beta(1 - \bar{q})) - \frac{\beta}{2}\bar{m}^2 + \frac{\alpha\beta\bar{q}}{2(1 - \beta(1 - \bar{q}))} \\ - \frac{\alpha\beta^2}{2}\frac{\bar{q}(1 - \bar{q})}{(1 - \beta(1 - \bar{q}))^2} + \mathbb{E}\ln\cosh\left(\beta z\sqrt{\frac{\alpha\bar{q}}{(1 - \beta(1 - \bar{q}))^2}} + \beta\bar{m}\right),\end{aligned} \tag{2.8}$$

where z is a random Gaussian variable with zero average and unit variance, $\mathbb{E}$ denotes the average over z and α is the load capacity of the network. In this limit, the order parameters $\bar{q}$ and $\bar{m}$ fulfill the celebrated Amit–Gutfreund–Sompolinsky self-consistency equations [20,21]:

$$\bar{q} = \mathbb{E}\tanh^2\left(\beta\bar{m} + \beta\sqrt{\frac{\alpha\bar{q}}{(1 - \beta(1 - \bar{q}))^2}}\, z\right), \tag{2.9}$$

$$\bar{m} = \mathbb{E}\tanh\left(\beta\bar{m} + \beta\sqrt{\frac{\alpha\bar{q}}{(1-\beta(1-\bar{q}))^2}}z\right). \tag{2.10}$$

On the other hand, within one step of the replica-symmetry breaking (1RSB) scheme [22–24] it is assumed that the distribution of the two-replica overlap q, in the thermodynamic limit, displays two delta-peaks at the equilibrium values $\bar{q}_0$ and $\bar{q}_1 > \bar{q}_0$ and the concentration on these two values is ruled by the parameter $\theta \in [0,1]$, while m self-averages as in the RS case :

$$\lim_{N\to+\infty} P_N(m) = \delta(m-\bar{m}_1), \tag{2.11}$$

$$\lim_{N\to+\infty} P'_N(q) = \theta\delta(q-\bar{q}_0) + (1-\theta)\delta(q-\bar{q}_1), \tag{2.12}$$

Within this assumption, the disorder-averaged free-energy is given by (see e.g. [24])

$$\begin{aligned} -\beta f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\beta,\alpha,\theta) = \ln 2 &- \frac{\alpha}{2}\ln\left(\Delta_1(\beta,\bar{q}_1)\right) + \frac{\alpha}{2\theta}\ln\left(\frac{\Delta_1(\beta,\bar{q}_1)}{\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1)}\right) + \frac{\alpha\beta\bar{q}_0}{2\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1)} \\ &- \frac{\beta}{2}\bar{m}_1^2 + \frac{1}{2}\alpha\beta^2\theta\frac{\bar{q}_0^2}{\Delta_2^2(\beta,\theta,\bar{q}_0,\bar{q}_1)} + \frac{1}{\theta}\mathbb{E}_1\ln\mathbb{E}_2\cosh^\theta g_\theta(\beta,\alpha,\bar{m}_1,\bar{q}_0,\bar{q}_1) \\ &- \frac{1}{2}\alpha\beta^2\theta\bar{q}_1\left(\frac{\bar{q}_0}{\Delta_2^2(\beta,\theta,\bar{q}_0,\bar{q}_1)} + \frac{\bar{q}_1-\bar{q}_0}{\Delta_1(\beta,\bar{q}_1)\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1)}\right) \\ &- \frac{1}{2}\alpha\beta^2(1-\bar{q}_1)\left(\frac{\bar{q}_0}{\Delta_2^2(\beta,\theta,\bar{q}_0,\bar{q}_1)} + \frac{\bar{q}_1-\bar{q}_0}{\Delta_1(\beta,\bar{q}_1)\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1)}\right) \end{aligned} \tag{2.13}$$

where, for mathematical convenience, we defined

$$\Delta_1(\beta,\bar{q}_1) := 1-\beta(1-\bar{q}_1) \tag{2.14}$$

$$\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1) := 1-\beta(1-\bar{q}_1) - \beta\theta(\bar{q}_1-\bar{q}_0) \tag{2.15}$$

$$g_\theta(\beta,\alpha,\bar{m}_1,\bar{q}_1,\bar{q}_0) := \beta\bar{m}_1 + \frac{\beta z^{(1)}\sqrt{\alpha\bar{q}_0}}{\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1)} + \beta z^{(2)}\sqrt{\frac{\alpha(\bar{q}_1-\bar{q}_0)}{\Delta_1(\beta,\bar{q}_1)\Delta_2(\beta,\theta,\bar{q}_0,\bar{q}_1)}} \tag{2.16}$$

and we have denoted with $\mathbb{E}_1$, $\mathbb{E}_2$ the averages w.r.t. the standard normal variables $z^{(1)}$ and $z^{(2)}$, respectively. From now on, we imply the dependence of all the functions on β and α. By extremizing the 1RSB free-energy w.r.t. its order parameters $\bar{q}_0$, $\bar{q}_1$, $\bar{m}_1$, it is possible to show that the latter fulfill the following self-consistency equations

$$\begin{aligned} \bar{m}_1 &= \mathbb{E}_1\left[\frac{\mathbb{E}_2\cosh^\theta g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)\tanh g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)}{\mathbb{E}_2\cosh^\theta g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)}\right], \\ \bar{q}_1 &= \mathbb{E}_1\left[\frac{\mathbb{E}_2\cosh^\theta g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)\tanh^2 g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)}{\mathbb{E}_2\cosh^\theta g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)}\right], \\ \bar{q}_0 &= \mathbb{E}_1\left[\frac{\mathbb{E}_2\cosh^\theta g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)\tanh g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)}{\mathbb{E}_2\cosh^\theta g_\theta(\bar{m}_1,\bar{q}_1,\bar{q}_0)}\right]^2. \end{aligned} \tag{2.17}$$

The key idea of our method is to assume that at the onset of the RS instability, one of the two delta-peaks in Eq. (2.12) has vanishing weight, i.e. either $\theta \to 0$ or $\theta \to 1$. Consistency with the RS theory then requires the dominating peak to be located at the value $\bar{q}$ of the RS order parameter, so either $\lim_{\theta\to 0}\bar{q}_1 = \bar{q}$ or $\lim_{\theta\to 1}\bar{q}_0 = \bar{q}$. As we will see below, both these relations are generally satisfied, hence we appeal to the physical interpretation of RS breaking to determine which scenario applies. Noting that when the RS theory becomes unstable, a multiplicity of states emerges with mutual overlap $\bar{q}_0$ and self-overlap $\bar{q}_1$, within the single state (or each of the states) predicted by the RS theory, it is natural to assume that the value of the mutual overlap between the newly born states is equal to the self-overlap $\bar{q}$ of the state(s) assumed by the RS theory. Thus, we will assume that at the onset of the RS instability, $\bar{q}_0 \to \bar{q}$, hence $\theta \to 1$. For the Hopfield model, taking this limit in (2.17), and using

$$g_1(\bar{m}_1,\bar{q}_1,\bar{q}_0) = \beta\bar{m}_1 + \beta z^{(1)}\frac{\sqrt{\alpha\bar{q}_0}}{\Delta_1(\bar{q}_0)} + \beta z^{(2)}\sqrt{\frac{\alpha(\bar{q}_1-\bar{q}_0)}{\Delta_1(\bar{q}_1)\Delta_1(\bar{q}_0)}}, \tag{2.18}$$

we get

$$\begin{aligned} \lim_{\theta\to 1}\bar{q}_0 &= \mathbb{E}_1\left[\frac{\mathbb{E}_2\sinh g_1(\bar{m}_1,\bar{q}_1,\bar{q}_0)}{\mathbb{E}_2\cosh g_1(\bar{m}_1,\bar{q}_1,\bar{q}_0)}\right]^2 \\ &= \mathbb{E}_1\left[\frac{\exp\left(\frac{\beta^2\alpha(\bar{q}_1-\bar{q}_0)}{2\Delta_1(\bar{q}_1)\Delta_1(\bar{q}_0)}\right)\sinh\left(\beta\bar{m}_1 + \beta\sqrt{\frac{\alpha\bar{q}_0}{(1-\beta(1-\bar{q}_0))^2}}z^{(1)}\right)}{\exp\left(\frac{\beta^2\alpha(\bar{q}_1-\bar{q}_0)}{2\Delta_1(\bar{q}_1)\Delta_1(\bar{q}_0)}\right)\cosh\left(\beta\bar{m}_1 + \beta\sqrt{\frac{\alpha\bar{q}_0}{(1-\beta(1-\bar{q}_0))^2}}z^{(1)}\right)}\right]^2 \end{aligned}$$

$$= \mathbb{E}_1 \tanh^2\left(\beta\bar{m}_1 + \beta\sqrt{\frac{\alpha\bar{q}_0}{(1-\beta(1-\bar{q}_0))^2}}\, z^{(1)}\right) \tag{2.19}$$

where in the second line we have used $\sinh(A+B) = \sinh A \cosh B + \sinh B \cosh A$ and $\cosh(A+B) = \cosh A \cosh B + \sinh B \sinh A$ (A denoting the first two terms on the RHS of (2.18) and B the last one) and have performed the integral over $z^{(2)}$ using the oddity of the sinh function. As (2.19) is identical to (2.9), in the limit $\theta \to 1$, $\bar{q}_0$ is indeed equal to the RS order parameter $\bar{q}$, as anticipated. Similarly, we can show that

$$\lim_{\theta\to 1} \bar{m}_1 = \bar{m} \tag{2.20}$$

and one can easily verify that $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)|_{\theta=1} = f_{RS}(\bar{m}, \bar{q})$, as expected from the fact that, for $\theta = 1$, Eq. (2.12) reduces to (2.7) and one retrieves the RS scheme. Our purpose is then to prove that for values of θ close but away from one, the 1RSB expression of the quenched free-energy is smaller than the RS expression, i.e. $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta) < f_{RS}(\bar{m}, \bar{q})$, below a critical line in the parameters space (α, β).

To this purpose, we expand the 1RSB quenched free-energy around $\theta = 1$ (i.e. around the replica symmetric expression) to the first order, writing

$$f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta) = f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)|_{\theta=1} + (\theta-1)\partial_\theta f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)|_{\theta=1}, \tag{2.21}$$

where $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)|_{\theta=1} = f_{RS}(\bar{m}, \bar{q})$. To determine when the RS solution becomes unstable, i.e. $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta) < f_{RS}(\bar{m}, \bar{q})$ we inspect the sign of $\partial_\theta f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)|_{\theta=1}$, keeping in mind that $\theta - 1 < 0$. To evaluate the latter, we need to expand the self-consistency equations for $\bar{m}_1$, $\bar{q}_0$ and $\bar{q}_1$ around $\theta = 1$, to linear orders in $\theta - 1$. We obtain

$$\bar{q}_0 = \mathbb{E}_1 \tanh^2\left(\beta\bar{m} + \beta\sqrt{\frac{\alpha\bar{q}_0}{(1-\beta(1-\bar{q}_0))^2}}\, z^{(1)}\right) + (\theta-1)A(\bar{m}_1, \bar{q}_0, \bar{q}_1) \tag{2.22}$$

where $A(\bar{m}_1, \bar{q}_0, \bar{q}_1)$ is a function of $\bar{m}_1$, $\bar{q}_0$ and $\bar{q}_1$ that will drop out of the calculation, whose expression is provided in (C.1). As to $\mathcal{O}((\theta-1)^0)$, $\bar{q}_0$ and $\bar{m}_1$ are equal to the RS order parameters $\bar{q}$ and $\bar{m}$, respectively, we can rewrite (2.22) as

$$\bar{q}_0 = \bar{q} + (\theta-1)A(\bar{m}, \bar{q}, \bar{q}_1). \tag{2.23}$$

Following the same path for $\bar{q}_1$, and using (2.23), we have

$$\bar{q}_1 =: \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_1, \bar{q}) \tanh^2 g_1(\bar{m}, \bar{q}_1, \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_1, \bar{q})}\right] + (\theta-1)B(\bar{m}_1, \bar{q}_0, \bar{q}_1) \tag{2.24}$$

where $B(\bar{m}_1, \bar{q}_0, \bar{q}_1)$ is provided in (C.2) and will drop out of the calculation. For $\theta = 1$, we have

$$\bar{q}_1 = \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_1, \bar{q}) \tanh^2 g_1(\bar{m}, \bar{q}_1, \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_1, \bar{q})}\right] \tag{2.25}$$

which is a self-consistency equation for $\bar{q}_1$, that depends only on $\bar{q}$ and $\bar{m}$. Denoting with $\tilde{q}_1(\bar{m}, \bar{q})$ its solution, we can then write (2.24) as

$$\bar{q}_1 = \tilde{q}_1(\bar{m}, \bar{q}) + (\theta-1)B(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q})) \tag{2.26}$$

and, finally,

$$\bar{q}_0 = \bar{q} + (\theta-1)A(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q})). \tag{2.27}$$

Similarly to $\bar{q}_0$ and $\bar{q}_1$ we can expand also $\bar{m}_1$ as

$$\bar{m}_1 = \bar{m} + (\theta-1)C(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q})) \tag{2.28}$$

where $C(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))$ is provided in (C.3).

Using (2.28), (2.26) and (2.27) to evaluate the derivative of $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)$ w.r.t. θ and finally setting $\theta = 1$, we obtain:

$$\begin{aligned} K(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q})) &:= \partial_\theta(-\beta f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta))|_{\theta=1} \\ &= -\frac{\alpha}{2}\log\left[\frac{\Delta_1(\tilde{q}_1(\bar{m}, \bar{q}))}{\Delta_1(\bar{q})}\right] + \frac{\alpha\beta(\tilde{q}_1(\bar{m}, \bar{q}) - \bar{q})}{2\Delta_1(\bar{q})} - \frac{\alpha\beta^2(1-\tilde{q}_1(\bar{m}, \bar{q}))(\tilde{q}_1(\bar{m}, \bar{q}) - \bar{q})}{2\Delta_1(\tilde{q}_1(\bar{m}, \bar{q}))\Delta_1(\bar{q})} \\ &\quad - \mathbb{E}\log\cosh\left(\beta\bar{m} + \beta z\frac{\sqrt{\alpha\bar{q}}}{2\Delta_1(\bar{q})}\right) + \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}) \log\cosh g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}\right]. \end{aligned} \tag{2.29}$$

Next, we study the sign of (2.29), where $\bar{q}$ and $\tilde{q}_1(\bar{m}, \bar{q})$ are the solutions of the self-consistency equations (2.9) and (2.25), respectively. To this purpose, it is useful to study the behavior of the function $K(\bar{m}, \bar{q}, x)$ for $x \in [0, \bar{q}]$. For $x = \bar{q}$, regardless of the value assigned to $\bar{m}$, we have $K(\bar{m}, \bar{q}, \bar{q}) = 0$, while the extremum of $K(\bar{m}, \bar{q}, x)$ is found from

$$\partial_x K(\bar{m}, \bar{q}, x) = \frac{\beta^2\alpha x}{2\Delta_1(x)^2}\left[x - \mathbb{E}_1\frac{\mathbb{E}_2 \cosh g_1(\bar{m}, x, \bar{q}) \tanh^2 g_1(\bar{m}, x, \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, x, \bar{q})}\right] = 0$$

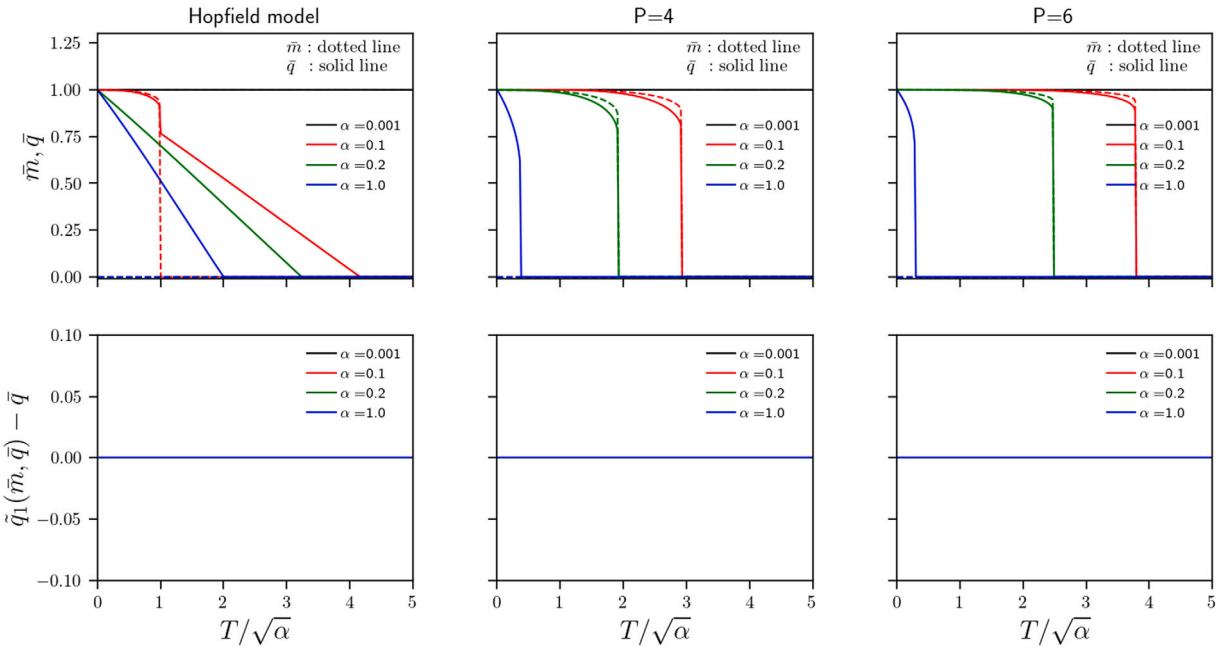

Fig. 1. RS overlap $\bar{q}$ (top row) and the difference between $\tilde{q}_1(\bar{m}, \bar{q})$ and $\bar{q}$ (bottom row) versus the scaled parameter $T/\sqrt{\alpha}$, for different values of α (as shown in the legend), for the Hopfield model (left) and Hebbian networks with P-node interactions, for $P = 4$ (mid) and $P = 6$ (right).

at

$$x = \mathbb{E}_1 \frac{\mathbb{E}_2 \cosh g_1(\bar{m}, x, \bar{q}) \tanh^2 g_1(\bar{m}, x, \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, x, \bar{q})} \equiv \tilde{q}_1(\bar{m}, \bar{q}), \tag{2.30}$$

where the last equality follows from Eq. (2.25). Given that $K(\bar{m}, \bar{q}, x)$ vanishes for $x = \bar{q}$, if the extremum $x = \tilde{q}_1(\bar{m}, \bar{q})$ is global in the domain considered, we must have that $K(\bar{q}, \tilde{q}_1(\bar{q})) > 0$ if $x = \tilde{q}_1(\bar{q})$ is a maximum and $K(\bar{m}, \bar{q}, \tilde{q}_1(\bar{q})) < 0$ if $x = \tilde{q}_1(\bar{m}, \bar{q})$ is a minimum. Therefore, if

$$\partial_x^2 K(\bar{q}, x)|_{x=\tilde{q}_1(\bar{m},\bar{q})} = -\frac{\beta^2\alpha}{2\Delta_1(\tilde{q}_1(\bar{m}, \bar{q}))^2} \left\{ 1 - \frac{\beta^2\alpha}{\Delta_1(\tilde{q}_1(\bar{m}, \bar{q}))^2} \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \operatorname{sech}^3 g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})} \right]^2 \right\} \tag{2.31}$$

is positive, $K(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))$ is negative and

$$f_{1RSB}(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}|\theta) = f_{RS}(\bar{m}, \bar{q}) - (\theta - 1)\frac{K(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}{\beta} < f_{RS}(\bar{m}, \bar{q}), \tag{2.32}$$

hence the RS theory becomes unstable when the expression in the curly brackets in (2.31) becomes negative i.e. for

$$(1 - \beta(1 - \tilde{q}_1(\bar{m}, \bar{q})))^2 < \beta^2\alpha\mathbb{E}_1 \left[\frac{\mathbb{E}_2 \operatorname{sech}^3 g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g_1(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})} \right]^2 \tag{2.33}$$

This expression recovers the result found by Coolen in [17] using the de Almeida–Thouless approach [6], in the limit $\tilde{q}_1(\bar{m}, \bar{q}) \to \bar{q}$, where (2.33) reduces to

$$(1 - \beta(1 - \bar{q}))^2 < \alpha\beta^2 \mathbb{E} \cosh^{-4} \left[\beta\bar{m} + \beta z \frac{\sqrt{\alpha\bar{q}}}{1 - \beta(1 - \bar{q})} \right]. \tag{2.34}$$

While this limit is *a priori* unjustified, as $\bar{q}$ and $\tilde{q}_1(\bar{m}, \bar{q})$ should be solved from the self-consistency equations (2.9) and (2.25), respectively, one can check numerically that the solutions of these equations are virtually indistinguishable for any temperature (see Fig. 1, left panel), and the resulting RS instability line is almost identical to the AT line derived in [17], see Fig. 2 (left panel). Small deviations can be appreciated in the retrieval region (see right panel), but these are likely due to numerical precision.

As our method does not rely on the assumption $\tilde{q}_1(\bar{m}, \bar{q}) \to \bar{q}$, it provides a more general approach than the one originally devised by de Almeida and Thouless, that can be carried over to systems with discontinuous phase transitions, where $\bar{q}_1$ differs from $\bar{q}_0$ even at the onset of the RS instability, implying $\tilde{q}_1(\bar{m}, \bar{q}) \neq \bar{q}$.

Before concluding this section, we note that, although we have disregarded the limit $\theta \to 0$ as lacking physical interpretation, such limit is still well defined mathematically and one may ask what would be the outcome of a similar analysis carried out in this

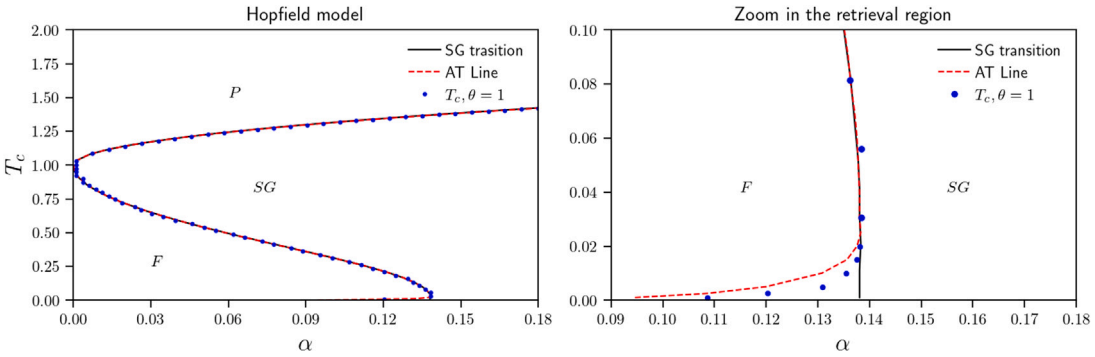

Fig. 2. RS instability line (i.e. T_c versus α) obtained via our method (blue crosses) and the AT line obtained in the limit $\bar{q}_1(\bar{m}, \bar{q}) \to \bar{q}$ (red dashed curve), for the Hopfield model. The several branches separate the paramagnetic (P), spin-glass (SG) and retrieval (M) region. The black curve show the critical temperature $T^\star$ at which $\bar{q}$ becomes non-zero within the RS theory, i.e. the spin-glass (SG) transition.

limit. We will perform such analysis in Appendix B. Intriguingly, we find that the analysis for $\theta \to 0$ gives the same instability line as the analysis for $\theta \to 1$, in all the models we considered, except in the spherical P-spin model. Reassuringly, in the latter case, the analysis at $\theta \to 0$ leads to a lower temperature for the RS instability, confirming that the limit $\theta \to 1$ gives the physical transition and it is therefore the relevant one.

3. Hebbian networks with P-node interactions

In this section we consider generalizations of the Hopfield model, where neurons interact in P-tuples of even $P \geq 4$ (rather than pairwise, i.e. $P = 2$). Such networks were shown to store many more patterns than the number of their nodes, so they work as *dense* associative memories [25]. They are also known to be dual to deep neural networks [1,26] and to exhibit information processing capabilities that are forbidden in shallow networks, such as the existence of a region in the parameter space where they can retrieve patterns although these are overshadowed by the noise [27]. As before, we consider a network of N interacting Ising neurons $\sigma_i \in \{1, -1\}$, with K stored patterns $\boldsymbol{\xi}^\mu$ and P-node interactions $J_{i_1 \dots i_P} = \frac{1}{P!} \sum_{\mu=1}^{K} \xi_{i_1}^\mu \dots \xi_{i_P}^\mu$. The Hamiltonian of this model can be written as

$$H_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{N^{1-P}}{P!} \sum_{\mu=1}^{K} \sum_{i_1,\dots,i_P} \xi_{i_1}^\mu \dots \xi_{i_P}^\mu \sigma_{i_1} \dots \sigma_{i_P} \tag{3.1}$$

The order parameters are still the Mattis magnetization m and the two-replicas overlap q, as introduced in (2.4), with their RS distributions given in (2.6) and (2.7) and their 1RSB generalizations given in (2.12) and (2.11). The quenched free-energy in RS assumption is (see [28])

$$\begin{aligned} -\beta' f_{RS}(\bar{m}, \bar{q}|\beta', \alpha) = \ln 2 - \frac{\beta'}{2}(P-1)\bar{m}^P + \frac{\beta'^2 \alpha}{4}(1 - \bar{q}^P) - \frac{\beta'^2 \alpha P}{4} \bar{q}^{P-1}(1 - \bar{q}) \\ + \mathbb{E} \ln \cosh \left(\beta' \frac{P}{2} \bar{m}^{P-1} + \beta' z \sqrt{\alpha \frac{P}{2} \bar{q}^{P-1}} \right) \end{aligned} \tag{3.2}$$

with $\beta' := 2\beta/P!$, where β is the inverse temperature and $\alpha = \lim_{N\to\infty} K/N^{P-1}$ is the network load. $\mathbb{E}$ is the average w.r.t. the standard Gaussian random variable z, and $\bar{q}$ and $\bar{m}$ satisfy the self-consistency equations:

$$\begin{aligned} \bar{m} &= \mathbb{E} \tanh \left(\beta' \frac{P}{2} \bar{m}^{P-1} + \beta' \sqrt{\alpha \frac{P}{2} \bar{q}^{P-1}} z \right), \\ \bar{q} &= \mathbb{E} \tanh^2 \left(\beta' \frac{P}{2} \bar{m}^{P-1} + \beta' \sqrt{\alpha \frac{P}{2} \bar{q}^{P-1}} z \right). \end{aligned} \tag{3.3}$$

On the other hand, the quenched free-energy within the 1RSB approximation (see [1]), reads as

$$\begin{aligned} -\beta' f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\beta', \alpha, \theta) = \ln 2 - \frac{\beta'}{2}(P-1)\bar{m}_1^P + \frac{\beta'^2 \alpha}{4} \left[1 - \theta \bar{q}_0^P + (\theta - 1)\bar{q}_1^P \right] \\ - \frac{\beta'^2 \alpha P}{4} \bar{q}_1^{P-1} - \frac{\beta'^2}{4} P \alpha \left[(\theta - 1)\bar{q}_1^P - \theta \bar{q}_0^P \right] \\ + \frac{1}{\theta} \mathbb{E}_1 \ln \mathbb{E}_2 \cosh^\theta g(\beta', \alpha, \bar{m}_1, \bar{q}_0, \bar{q}_1) \end{aligned} \tag{3.4}$$

where $\mathbb{E}_1$, $\mathbb{E}_2$ are the average w.r.t. the standard normal random variables $z^{(1)}$ and $z^{(2)}$, respectively, and

$$g(\bar{m}_1, \bar{q}_1, \bar{q}_0|\beta', \alpha) = \frac{\beta' P}{2}\bar{m}_1^{P-1} + \beta' z^{(1)}\sqrt{\frac{P}{2}\alpha\bar{q}_0^{P-1}} + \beta' z^{(2)}\sqrt{\frac{P}{2}\alpha\left(\bar{q}_1^{P-1} - \bar{q}_0^{P-1}\right)}. \tag{3.5}$$

In this approximation the self-consistency equations for the order parameter $\bar{q}_1$, $\bar{q}_0$ and $\bar{m}$ are as in (2.17), with the argument of the hyperbolic cosine and tangent replaced by (3.5). As before, for $\theta = 1$, $\bar{q}_0 = \bar{q}$ and the 1RSB expression for the quenched free-energy reduces to the RS one.

From now on, we imply the dependence of the functions on β' and α. Our objective is to prove that for θ close but away from one, the 1RSB quenched free-energy is smaller than its replica symmetric counterpart i.e. $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta) < f_{RS}(\bar{m}, \bar{q})$ above a critical value of the effective parameter $\sqrt{\alpha}\beta'$. To this purpose, we proceed as in the Hopfield model: we expand, to the leading order in $\theta - 1$, the 1RSB quenched free-energy around its RS expression, as shown in (2.21). Since the self-consistency equations also depend on θ, we need to expand them too. Following the same steps as in the Hopfield model, we can write $\bar{m}_1$ as in (2.28) with $C(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))$ as in (C.9), where $\tilde{q}_1(\bar{m}, \bar{q})$ is the solution of the self-consistency equation

$$\tilde{q}_1 = \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g(\bar{m}, \tilde{q}_1, \bar{q}) \tanh^2 g(\bar{m}, \tilde{q}_1, \bar{q})}{\mathbb{E}_2 \cosh g(\bar{m}, \tilde{q}_1, \bar{q})}\right], \tag{3.6}$$

$\bar{q}_0$ as in (2.27), with $A(\bar{m}_1, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))$ given in (C.7), and $\bar{q}_1$ as given in (2.26) with $B(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))$ given in (C.8). With the above expressions in hand, we can now calculate the derivative of $f_{1\text{RSB}}$ w.r.t. θ when $\theta = 1$, as needed in (2.21)

$$\begin{aligned} &K(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}) := \partial_\theta(-\beta' f_{1\text{RSB}}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta))|_{\theta=1} \\ &= -\frac{\beta'^2\alpha}{4}(P-1)[(\tilde{q}_1(\bar{m}, \bar{q}))^P - \bar{q}^P] - \frac{\beta'^2\alpha}{4}P[(\tilde{q}_1(\bar{m}, \bar{q}))^{P-1} - \bar{q}^{P-1}] \\ &- \mathbb{E}\ln\cosh\left(\beta'\sqrt{\frac{P}{2}\bar{q}^{P-1}}z + \beta'\bar{m}^{P-1}\right) + \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}) \log\cosh g(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}\right] \end{aligned} \tag{3.7}$$

Again, we have that $K(\bar{m}, \bar{q}, \bar{q}) = 0$, regardless of the value assigned to $\bar{m}$, (this follows from the fact that for $\theta = 1$, $\bar{q}$ is an extremum of the free-energy). Next, we inspect the sign of $K(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})$. To this purpose, we study $K(\bar{m}, \bar{q}, x)$ for $x \in [0, \bar{q}]$ and locate its extrema, which are found from

$$\partial_x K(\bar{m}, \bar{q}, x) = -\frac{\beta'^2\alpha}{4}(P-1)Px^{P-2}\left[x - \mathbb{E}_1\left[\frac{\mathbb{E}_2 \sinh g(\bar{m}, \bar{q}, x) \tanh g(\bar{m}, \bar{q}, x)}{\mathbb{E}_2 \cosh g(\bar{m}, \bar{q}, x)}\right]\right] = 0 \tag{3.8}$$

as

$$x = \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g(x, \bar{q}) \tanh^2 g(x, \bar{q})}{\mathbb{E}_2 \cosh g(x, \bar{q})}\right] \equiv \tilde{q}_1(\bar{m}, \bar{q}) \tag{3.9}$$

where the last equality follows from (3.6). Under the assumption that the extremum $x = \tilde{q}_1(\bar{m}, \bar{q})$ is global in the domain considered and reasoning as in the Hopfield case, we have that $K(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}) > 0$ if $x = \tilde{q}_1(\bar{m}, \bar{q})$ is a maximum and $K(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}) < 0$ if it is a minimum. In particular, if

$$\begin{aligned} \partial_x^2 K(\bar{m}, \bar{q}, x)|_{x=\tilde{q}_1(\bar{m}, \bar{q})} = &-\frac{\beta'^2\alpha(P-1)P}{4}(\tilde{q}_1(\bar{m}, \bar{q}))^{P-2} \\ &\cdot\left\{1 - \frac{\beta'^2\alpha}{2}(P-1)P(\tilde{q}_1(\bar{m}, \bar{q}))^{P-2}\mathbb{E}_1\left[\frac{\mathbb{E}_2\text{sech}^3 g(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q})}\right]\right\} \end{aligned} \tag{3.10}$$

is positive, $K(\bar{m}, \tilde{q}_1(\bar{m}, \bar{q}), \bar{q}) < 0$ and $f_{1RSB} < f_{RS}$. This happens when the expression in the curly brackets of the equation above is negative, i.e. when the parameter $\alpha\beta'^2$ satisfies the inequality

$$\frac{\beta'^2\alpha}{2}(P-1)P(\tilde{q}_1(\bar{m}, \bar{q}))^{P-2}\mathbb{E}_1\left[\frac{\mathbb{E}_2\text{sech}^3 g(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))}{\mathbb{E}_2 \cosh g(\bar{m}, \bar{q}, \tilde{q}_1(\bar{m}, \bar{q}))}\right] > 1. \tag{3.11}$$

As noted in [1,29], Hebbian networks with P-node interactions are equivalent to Ising P-spin models under a suitable rescaling of the temperature $\beta'\sqrt{\alpha} \to \beta'$. With this rescaling, (3.11) retrieves indeed the RS instability line of the Ising P-spin model, that we have for completeness derived in Appendix A, using our method (see Eq. (A.30)). In the limit $\tilde{q}_1(\bar{m}, \bar{q}) \to \bar{q}$, (A.30) retrieves the AT line of the Ising P-spin model [18]. In Fig. 1 (mid and right panels) we plot the difference between $\tilde{q}_1(\bar{m}, \bar{q})$ and $\bar{q}$ for Hebbian networks with P-node interactions (obtained solving numerically the self-consistency equations (3.6) and (3.3)) as a function of the scaled parameter $T/\sqrt{\alpha}$, for different values of α. As for the Hopfield model, we find that $\tilde{q}_1(\bar{m}, \bar{q})$ is indistinguishable from $\bar{q}$, hence the limit $\tilde{q}_1(\bar{m}, \bar{q}) \to \bar{q}$ can be justified *a posteriori*.

In Fig. 3 we show the RS instability lines resulting from our method and the classic AT line obtained in the limit $\tilde{q}_1(\bar{m}, \bar{q}) \to \bar{q}$, for different values of P. The two lines coincides for all values of P. As explained earlier, we could have expanded the free-energy around $\theta = 0$ (as opposed to $\theta = 1$). In Appendix B we show that such analysis leads to the same line.

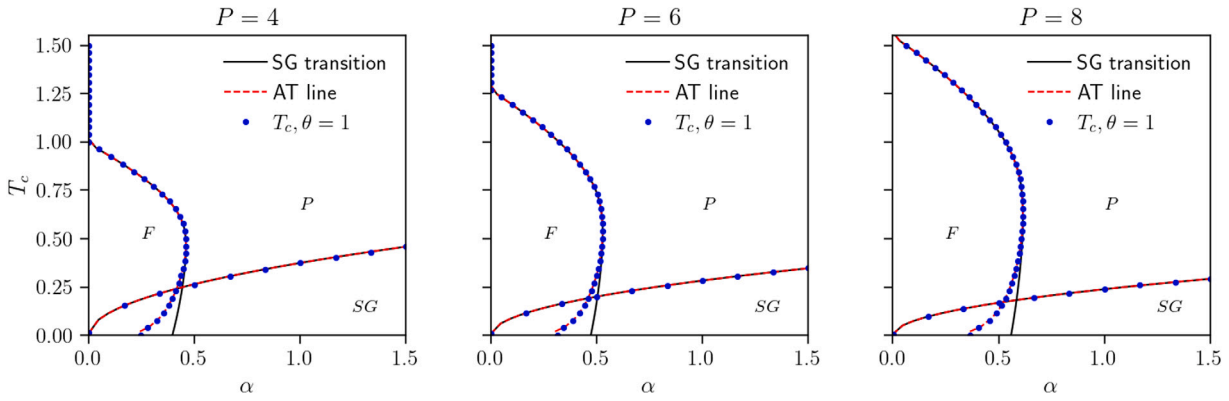

Fig. 3. RS instability line (i.e. T_c versus α) obtained via our method, i.e. by expanding the free-energy around $\theta = 1$ (blue dots) and the AT line obtained in the limit $\tilde{q}_1(\bar{m}, \bar{q}) \to \bar{q}$ (red curve), for the Hebbian network with P-node interactions, with $P = 2, 4, 6$ from left to right. The black curves show the critical temperature $T^\star$ at which $\bar{q}$ becomes non-zero within the RS theory, i.e. the spin-glass (SG) transition.

4. Discussion

In this work we proposed a simple and systematic method to derive the critical line in the parameter space (α, β), below which the 1RSB expression for the free-energy is smaller than the RS expression, in Hebbian neural networks. The same analysis for spin-glass models is carried out in Appendix A. For the Hopfield model, our approach recovers the critical line obtained by Coolen using the AT approach [17] as a special limit. Similarly, we recover the known AT lines of all the spin-glass models considered in the appendix, in the same limit, showing that our method provides a generalization of the approach originally devised by de Almeida and Thouless. Owing to its simplicity, our method allows for straightforward application to Hebbian networks with multi-node interactions, for which the AT-line was unknown.

The key idea of our method is to regard the 1RSB theory, which assumes two delta-peaks in the overlap distribution $P(q)$, located at $\bar{q}_1$ and $\bar{q}_0 < \bar{q}_1$, with weights $1 - \theta$ and θ, respectively, as departing continuously from the RS theory, which assumes only one peak at $\bar{q}$. This leads us to assume that at the onset of the RS instability, the 1RSB overlap distribution is dominated by one peak, so that either $\theta \to 0$ or $\theta \to 1$. Then, consistency with the RS theory requires either $\lim_{\theta \to 0} \bar{q}_1 = \bar{q}$ or $\lim_{\theta \to 1} \bar{q}_0 = \bar{q}$. As the physical interpretation of RS breaking suggests that the mutual overlap between 1RSB states should be equal to the self-overlap of the RS states, we regard the 1RSB theory as a continuous variation of the RS theory, when the parameter θ is decreased from one. Crucially, we do not make any assumption on the location of the peaks of the 1RSB theory, which are fixed by the 1RSB self-consistency equations. We then compare the 1RSB and the RS free-energies when θ (the only free-parameter in our analysis) is close to one, by performing simple expansions to linear orders in $\theta - 1$. In doing so, we solely require that $f_{1RSB}, \bar{q}_1, \bar{q}_0$ and $\bar{m}_1$ are differentiable up to the first order in a neighborhood of $\theta = 1$ and that the derivative of f_{1RSB} exists at $\theta = 1$.

Although our method is similar in spirit to the one introduced by Toninelli in [15], there is a crucial difference, in that the latter relies on the assumption $\bar{q}_1 \to \bar{q}$, which is, in our view, unjustified a priori. In fact, while $\bar{q}_0 = \bar{q}$ for $\theta = 1$, $\bar{q}_1$ may differ from $\bar{q}$, even in the limit $\theta \to 1$. This consideration also leads to a departure of our approach from the method originally devised by de Almeida and Thouless, which relies on a variation of the RS free-energy as the *order parameters* are varied continuously around their RS values. In contrast, we study the variation of the RS free-energy as the *statistical weight* of the order parameters is varied continuously (rather than the actual value of the order parameters). This approach allows us to determine the instability line of the RS theory also in spin-glass models which exhibit a discontinuous phase transition. As a prototypical example of this class of models, we consider in Appendix A the spherical P-spin model [22].

In conclusion, in this work we have presented a new method to find the instability line of the RS approximation. A compelling advantage of our method, when compared to the approach by de Almeida and Thouless [6], is that it does not require to compute the full eigenspectrum of the Hessian of the quadratic fluctuations of the free-energy around its RS value and it does not rely on the availability of an "ansatz-free" expression for the free-energy. This makes the computations easier and affordable also for neural networks. The method still requires the availability of explicit expressions for the RS and 1RSB free-energies, which however can be computed using different techniques (e.g. Guerra's interpolation) in addition to the replica trick. Our approach can in principle be extended to compute the stability of the k-RSB solution, expanding the corresponding $k + 1$-RSB free-energy, provided one has an explicit expression for the two. For example, it would be interesting to see whether the well-known transition from 1RSB to full-RSB occurring at the Gardner temperature [18] in the Ising P-spin model can be recovered within our approach, by studying the stability of the 2RSB solution. In addition, in recent years there has been a boost of renewed interest in mixed P-spherical models, as introduced in [30,31], as they have been shown to display unexpected dynamical behavior [32] and new types of spin-glass phases [33,34] as well as for their relevance to the modeling of random lasers [35–39]. These models similarly display transitions from 1RSB to full-RSB and would be good lab systems to test extensions of our theory. Another interesting avenue for future work would be a generalization of the 1-RSB scheme, which does not assume the Mattis magnetization to be self-averaging. Preliminary work in [24] suggests that the RS assumption is not the right approximation for Mattis magnetization.

Finally, an attractive perspective would be to apply our approach to predict the onset of ergodicity breaking in systems with *sparse* interactions. In such systems, the free-energy is typically expressed, already at the simplest RS level, in terms of order-parameter *functions* to be determined self-consistently. At 1RSB level, recursive equations for functional distributions of such functions must be solved. Working out the fluctuations of ansatz-free free-energies around their RS value, as it would be required by the de Almeida and Thouless approach would be unfeasible and no similar approach has been proposed to date. An alternative approach has been devised in [40], for the Bethe lattice with regular degrees, however it strongly relies on the assumptions of homogeneity in the network nodes and large degrees. We envisage that, owing to its simplicity, our method may carry over to more general sparse systems, where it would require an expansion of the functional 1RSB self-consistency equation to linear orders around the RS equation, which should be feasible.

CRediT authorship contribution statement

L. Albanese: Calculations provided, Writing of the paper. **A. Alessandrelli:** Calculations provided, Writing of the paper. **A. Annibale:** Guiding, Writing of the paper. **A. Barra:** Writing of the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported by Ministero degli Affari Esteri e della Cooperazione Internazionale (MAECI) via the BULBUL grant (Italy-Israel), CUP Project n. F85F21006230001.

L.A. acknowledges E. Zegna Founder's Scholarship, UMI (Unione Matematica Italiana), INDAM–GNFM Project (CUP E53C22001930001) and PRIN grant *Stochastic Methods for Complex Systems* n. 2017JFFHS for financial support and the Department of Mathematics at King's College London for the kind hospitality. L.A. and A.B. acknowledge INDAM (Istituto Nazionale d'Alta Matematica) and the PRIN grant *Statistical Mechanics of Learning Machines* n. 20229T9EAT for support. All the authors acknowledge the stimulating research environment provided by the Alan Turing Institute's Theory and Methods Challenge Fortnights event "Physics-informed Machine Learning".

Appendix A. Applications to Spin-glasses models

In this appendix, we derive the critical line for the instability of the RS theory for three archetypical spin-glass models, namely the SK model, the Ising P-spin and the spherical P-spin model, using the technique developed in the main text. In all cases, we will recover the AT lines known in the literature for the three models, in a specific limit, confirming the validity and higher generality of our approach.

A.1. Sherrington-Kirkpatrick model

The Sherrington-Kirkpatrick (SK) model [41] is a system of N Ising spins $\sigma_i \in \{\pm 1\}$ interacting via symmetric pairwise interactions J_{ij} which are i.i.d. Gaussian variables with zero average and variance J^2/N. The Hamiltonian of the model is

$$H_N(\boldsymbol{\sigma}|J) := -\frac{1}{2}\sum_{i,j}^{N,N} J_{ij}\sigma_i\sigma_j \tag{A.1}$$

and the order parameter is the two-replica overlap q as defined in (2.4). The quenched free-energy in RS assumption is [41]

$$-\beta f_{RS}(\bar{q}|\beta, J) = \ln 2 + \frac{\beta^2 J^2}{4}(1-\bar{q})^2 + \mathbb{E}\left[\ln\cosh\left(\beta J\sqrt{\bar{q}}z\right)\right] \tag{A.2}$$

where $\mathbb{E}$ is the average w.r.t. the Gaussian variable z and the order parameter $\bar{q}$ fulfills the self-consistency equation

$$\bar{q} = \mathbb{E}\tanh^2\left(\beta J\sqrt{\bar{q}}z\right). \tag{A.3}$$

The 1RSB approximation of the quenched free-energy is (see e.g. [42])

$$\begin{aligned}-\beta f_{1RSB}(\bar{q}_1, \bar{q}_0|\beta, J, \theta) = \ln 2 + \frac{\beta^2 J^2}{2}(1-\bar{q}_1) + \frac{1}{\theta}\mathbb{E}_1\left[\ln \mathbb{E}_2 \cosh^\theta g(\beta, J, \bar{q}_1, \bar{q}_0)\right] \\ - \frac{\beta^2 J^2}{4}\left[1 - \bar{q}_1^2 + \theta(\bar{q}_1^2 - \bar{q}_0^2)\right],\end{aligned} \tag{A.4}$$

where

$$g(\bar{q}_1, \bar{q}_0|\beta, J) = \beta J \sqrt{\bar{q}_1 - \bar{q}_0} z^{(2)} + \beta J \sqrt{\bar{q}_0} z^{(1)} \tag{A.5}$$

and $\mathbb{E}_1$, $\mathbb{E}_2$ are the average w.r.t. the standard Gaussian variables $z^{(1)}$ and $z^{(2)}$, respectively. From now on, we imply the dependence of g, f_{RS} and f_{1RSB} on β and J. The order parameters $\bar{q}_0$ and $\bar{q}_1$ fulfill the following self-consistency equations

$$\begin{aligned}
\bar{q}_1 &= \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \cosh^\theta g(\bar{q}_1, \bar{q}_0) \tanh^2 g(\bar{q}_1, \bar{q}_0)}{\mathbb{E}_2 \cosh^\theta g(\bar{q}_1, \bar{q}_0)} \right], \\
\bar{q}_0 &= \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \cosh^\theta g(\bar{q}_1, \bar{q}_0) \tanh g(\bar{q}_1, \bar{q}_0)}{\mathbb{E}_2 \cosh^\theta g(\bar{q}_1, \bar{q}_0)} \right]^2 .
\end{aligned} \tag{A.6}$$

Noting that $\lim_{\theta \to 1} \bar{q}_0 = \bar{q}$ and $f_{1RSB}(\bar{q}_1, \bar{q}_0|\theta = 1) = f_{RS}(\bar{q})$, our objective is to expand the 1RSB quenched free-energy for $\theta \simeq 1$. To this purpose, we expand the self-consistency equations for $\bar{q}_0$ and $\bar{q}_1$ to linear order in $\theta - 1$. Proceeding as in the Hopfield model, we get

$$\bar{q}_0 = \bar{q} + (\theta - 1) A(\bar{q}, \tilde{q}_1(\bar{q})) \tag{A.7}$$

$$\bar{q}_1 = \tilde{q}_1(\bar{q}) + (\theta - 1) B(\bar{q}, \tilde{q}_1(\bar{q})) \tag{A.8}$$

where $\tilde{q}_1(\bar{q})$ solves the self-consistency equation

$$\bar{q}_1 = \mathbb{E}_1 \left\{ \frac{\mathbb{E}_2 \cosh g(\bar{q}_1, \bar{q}) \tanh^2 g(\bar{q}_1, \bar{q})}{\mathbb{E}_2 \cosh g(\bar{q}_1, \bar{q})} \right\} \tag{A.9}$$

and $A(\bar{q}_0, \bar{q}_1)$ and $B(\bar{q}_0, \bar{q}_1)$ are given in (C.13) and (C.14), respectively. Next, we derive w.r.t. θ the 1RSB free-energy (A.4) where we replace $\bar{q}_0$ and $\bar{q}_1$ with (A.7), (A.8), obtaining

$$\begin{aligned}
\partial_\theta(-\beta f_{1RSB}(\bar{q}_1, \bar{q}_0, |\theta)) = &- \frac{\beta^2 J^2}{4} \left[(\tilde{q}_1(\bar{q}))^2 - \bar{q}^2 \right] - \frac{1}{\theta^2} \mathbb{E}_1 \ln \mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q}) \\
&- \frac{\beta^2 J^2}{2} B(\tilde{q}_1(\bar{q}), \bar{q}) \bar{q} + \frac{1}{\theta} \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q}) \log \cosh g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q})} \right] \\
&+ \frac{\beta^2 J^2}{2} B(\tilde{q}_1(\bar{q}), \bar{q}) \theta \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q}) \tanh g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q})} \right]^2 ,
\end{aligned} \tag{A.10}$$

which for $\theta = 1$, using (A.3) and similar manipulations to those used in (2.19), evaluates to

$$\begin{aligned}
K(\tilde{q}_1(\bar{q}), \bar{q}) := \partial_\theta(-\beta f_{1RSB}(\bar{q}_1, \bar{q}_0, |\theta))|_{\theta=1} = &- \frac{\beta^2 J^2}{4} \left[(\tilde{q}_1(\bar{q}))^2 - \bar{q}^2 \right] - \frac{\beta^2 J^2}{2} (\tilde{q}_1(\bar{q}) - \bar{q}) \\
&- \mathbb{E} \ln \cosh \left(\beta \sqrt{\bar{q}} z \right) + \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q}) \log \cosh g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q})} \right] .
\end{aligned} \tag{A.11}$$

We then study the sign of (A.11), where $\bar{q}$ and $\tilde{q}_1(\bar{q})$ are the solutions of the self-consistency equations (A.3) and (A.9), respectively. To this purpose, it is useful to study the behavior of the function $K(\bar{q}, x)$ for $x \in [0, \bar{q}]$. For $x = \bar{q}$, we have $K(\bar{q}, \bar{q}) = 0$, while the extremum of $K(\bar{q}, x)$ is found from

$$\partial_x K(x, \bar{q}) = -\frac{\beta^2 J^2}{2} x + \frac{\beta^2 J^2}{2} \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \sinh g(x, \bar{q}) \tanh g(x, \bar{q})}{\mathbb{E}_2 \cosh g(x, \bar{q})} \right] = 0, \tag{A.12}$$

as

$$x = \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \sinh g(x, \bar{q}) \tanh g(x, \bar{q})}{\mathbb{E}_2 \cosh g(x, \bar{q})} \right] \equiv \tilde{q}_1(\bar{q}) \tag{A.13}$$

via Eq. (A.9). Given that $K(\bar{q}, x)$ vanishes for $x = \bar{q}$, if the extremum $x = \tilde{q}_1(\bar{q})$ is global in the domain considered, we must have that $K(\tilde{q}_1(\bar{q}), \bar{q}) > 0$ if $x = \tilde{q}_1(\bar{q})$ is a maximum and $K(\tilde{q}_1(\bar{q}), \bar{q}) < 0$ if $x = \tilde{q}_1(\bar{q})$ is a minimum. Therefore, if

$$\partial_x^2 K(x, \bar{q})|_{x=\tilde{q}_1(\bar{q})} = -\frac{\beta^2 J^2}{2} \left(1 - \frac{\beta^2 J^2}{2} \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \text{sech}^3 g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q})} \right] \right) \tag{A.14}$$

is positive, $K(\bar{q}, \tilde{q}_1(\bar{q}))$ is negative, so the RS theory becomes unstable for

$$1 - \frac{\beta^2 J^2}{2} \mathbb{E}_1 \left[\frac{\mathbb{E}_2 \text{sech}^3 \, g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q})} \right] < 0. \tag{A.15}$$

We note that for $\tilde{q}_1(\bar{q}) \to \bar{q}$, such condition retrieves the well-known AT line [6]

$$1 - \beta^2 J^2 \mathbb{E} \left(\frac{1}{\cosh^4 \left(\beta J \sqrt{\bar{q}} z \right)} \right) < 0. \tag{A.16}$$

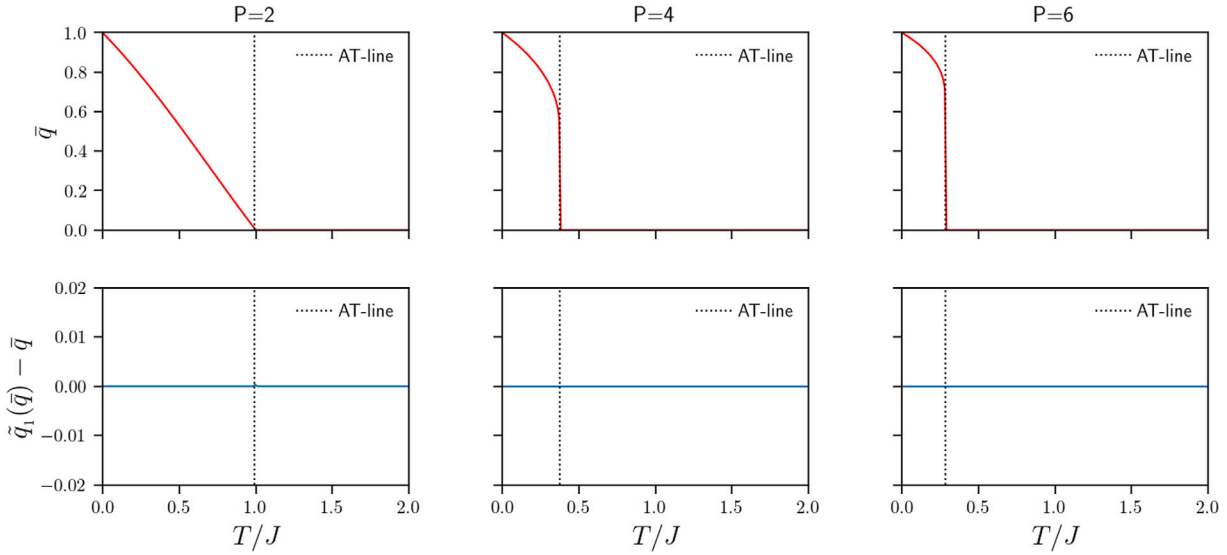

Fig. 4. RS overlap $\bar{q}$ (top) and difference between $\tilde{q}_1(\bar{q})$ and $\bar{q}$ (bottom) versus the ratio T/J, for the SK model (left) and the Ising P-spin model, with $P = 4$ (mid) and $P = 6$ (right panel). The dotted vertical line marks the onset of the RS instability.

By numerically solving the self-consistency equations (A.3) and (A.9), we can verify that $\tilde{q}_1(\bar{q}) = \bar{q}$ for all temperatures, hence this limit can be justified *a posteriori* for the SK model (see Fig. 4, left panel). We anticipate that this remains true for the Ising P-spin model analyzed in the next section.

A.2. The Ising P-spin model

In this section, we consider a system of N Ising spins $\sigma_i = \pm 1$, $i = 1, \dots, N$ governed by the Hamiltonian

$$H_N(\boldsymbol{\sigma}|J) = -\frac{1}{N^{P-1}P!}\sum_{i_1,\dots,i_P=1,\dots,1}^{N,\dots,N} J_{i_1,\dots,i_P}\sigma_{i_1}\dots\sigma_{i_P} \tag{A.17}$$

where $J_{i_1,\dots,i_P}$ are Gaussian i.i.d. variables, $J_{i_1,\dots,i_P} \sim \mathcal{N}(0, J^2)$. As in the previous cases, the order parameter of the model is the two-replica overlap q defined in (2.4). The quenched free-energy in RS assumption, at inverse temperature β reads as [18]

$$-\beta' f_{RS}(\bar{q}|\beta', J) = \ln 2 + \frac{\beta'^2 J^2}{4}[1 - P\bar{q}^{P-1} + (P-1)\bar{q}^P] + \mathbb{E}\ln\cosh\left(\beta' J z\sqrt{\frac{P}{2}\bar{q}^{P-1}}\right) \tag{A.18}$$

where $\beta' = 2\beta/P!$, $\mathbb{E}$ is the average w.r.t. the Gaussian variable z and $\bar{q}$ fulfills the following self-consistency equation:

$$\bar{q} = \mathbb{E}\tanh^2\left(\beta' J z\sqrt{\frac{P}{2}\bar{q}^{P-1}}\right). \tag{A.19}$$

The 1RSB expression of the quenched free-energy in the thermodynamic limit is [18]

$$\begin{aligned}-\beta' f_{1RSB}(\bar{q}_1, \bar{q}_0|\beta', J, \theta) = \ln 2 &+ \frac{\beta'^2 J^2}{4}\left[1 - P\bar{q}_1^{P-1} + (P-1)\bar{q}_1^P\right] \\ &+ \frac{1}{\theta}\mathbb{E}_1 \ln \mathbb{E}_2 \cosh^\theta g(\beta', J, \bar{q}_1, \bar{q}_0) - \frac{\beta'^2 J^2}{4}(P-1)\theta(\bar{q}_1^P - \bar{q}_0^P)\end{aligned} \tag{A.20}$$

where

$$g(\beta', J, \bar{q}_1, \bar{q}_0) = \beta' J z^{(1)}\sqrt{\frac{P}{2}\bar{q}_0^{P-1}} + \beta' J z^{(2)}\sqrt{\frac{P}{2}(\bar{q}_1^{P-1} - \bar{q}_0^{P-1})}, \tag{A.21}$$

$\mathbb{E}_1$, $\mathbb{E}_2$ are the average w.r.t. the standard Gaussian variables $z^{(1)}$ and $z^{(2)}$ and $\bar{q}_0$, $\bar{q}_1$ fulfill the following self-consistency equations

$$\begin{aligned}\bar{q}_1 &= \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh^\theta g(\beta', J, \bar{q}_1, \bar{q}_0)\tanh^2 g(\beta', J, \bar{q}_1, \bar{q}_0)}{\mathbb{E}_2 \cosh^\theta g(\beta', J, \bar{q}_1, \bar{q}_0)}\right], \\ \bar{q}_0 &= \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh^\theta g(\beta', J, \bar{q}_1, \bar{q}_0)\tanh(\beta', J, \bar{q}_1, \bar{q}_0)}{\mathbb{E}_2 \cosh^\theta g(\beta', J, \bar{q}_1, \bar{q}_0)}\right]^2.\end{aligned} \tag{A.22}$$

From now on, we imply the dependence of g and of the RS and 1RSB quenched free energies f_{RS}, f_{1RSB} on β' and J. The aim is to prove that the 1RSB approximation of the quenched free-energy is smaller than the replica symmetric one, above a critical value of the parameter $\beta' J$.

As before, we note that $\lim_{\theta\to 1}\bar{q}_0 = \bar{q}$ and $f_{1RSB}(\bar{q}_1, \bar{q}_0|\theta = 1) = f_{RS}(\bar{q})$ and we aim at expanding the 1RSB quenched free-energy for $\theta \simeq 1$. To this purpose, we first expand the self-consistency equations for $\bar{q}_0$ and $\bar{q}_1$ to linear order in $\theta - 1$, to obtain

$$\bar{q}_0 = \bar{q} + (\theta - 1)A(\bar{q}, \tilde{q}_1(\bar{q})) \tag{A.23}$$

$$\bar{q}_1 = \tilde{q}_1(\bar{q}) + (\theta - 1)B(\bar{q}, \tilde{q}_1(\bar{q})) \tag{A.24}$$

where $\tilde{q}_1(\bar{q})$ solves the self-consistency equation

$$\bar{q}_1 = \mathbb{E}_1\left\{\frac{\mathbb{E}_2 \cosh g(\bar{q}_1, \bar{q}) \tanh^2 g(\bar{q}_1, \bar{q})}{\mathbb{E}_2 \cosh g(\bar{q}_1, \bar{q})}\right\} \tag{A.25}$$

and $A(\bar{q}_0, \bar{q}_1)$ and $B(\bar{q}_0, \bar{q}_1)$ are given (C.15) and (C.16), respectively. Then, we derive w.r.t. θ the 1RSB free-energy (A.20) where we replace $\bar{q}_0$ and $\bar{q}_1$ with (A.23), (A.24), obtaining

$$\begin{aligned}\partial_\theta(-\beta' f_{1RSB}(\bar{q}_1, \bar{q}_0, |\theta)) = &-\frac{\beta'^2 J^2}{4}(P-1)((\tilde{q}_1(\bar{q}))^P - \bar{q}^P) - \frac{1}{\theta^2}\mathbb{E}_1 \ln \mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q}) \\ &+ \frac{1}{\theta}\mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q}) \log\cosh g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q})}\right] \\ &+ \frac{\beta'^2 J^2}{4}P(P-1)B(\tilde{q}_1(\bar{q}), \bar{q})\theta\bar{q}^{P-2}\mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q}) \tanh g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh^\theta g(\tilde{q}_1(\bar{q}), \bar{q})}\right]^2 \\ &- \frac{\beta'^2 J^2}{4}P(P-1)B(\tilde{q}_1(\bar{q}), \bar{q})\bar{q}^{P-1},\end{aligned} \tag{A.26}$$

which, for $\theta = 1$, using (A.19) and performing similar calculations to those in (2.19), evaluates to

$$\begin{aligned}K(\tilde{q}_1(\bar{q}), \bar{q}) := \partial_\theta(-\beta' f_{1RSB}(\bar{q}_1, \bar{q}_0, |\theta))|_{\theta=1} = &-\frac{\beta'^2 J^2}{4}(P-1)((\tilde{q}_1(\bar{q}))^P - \bar{q}^P) \\ &+ \mathbb{E}_1\left[\frac{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q}) \log\cosh g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q})}\right] \\ &- \mathbb{E}\ln\cosh\left(\beta'\sqrt{\frac{P}{2}\bar{q}^{P-1}}\right) - \frac{\beta'^2 J^2}{4}P((\tilde{q}_1(\bar{q}))^{P-1} - \bar{q}^{P-1}).\end{aligned} \tag{A.27}$$

Next, we study the sign of (A.27), where $\bar{q}$ and $\tilde{q}_1(\bar{q})$ are the solutions of the self-consistency equations (A.19) and (A.25), respectively. To this purpose, we study the behavior of the function $K(x, \bar{q})$ for $x \in [0, \bar{q}]$. For $x = \bar{q}$, we have $K(\bar{q}, \bar{q}) = 0$, while the extremum of $K(\bar{q}, x)$ is found from

$$\partial_x K(x, \bar{q}) = -\frac{\beta'^2 J^2}{4}(P-1)Px^{P-1} + \frac{\beta'^2 J^2}{4}(P-1)Px^{P-2}\mathbb{E}_1\left[\frac{\mathbb{E}_2 \sinh g(x, \bar{q}) \tanh g(x, \bar{q})}{\mathbb{E}_2 \cosh g(x, \bar{q})}\right] = 0, \tag{A.28}$$

as

$$x = \mathbb{E}_1\left[\frac{\mathbb{E}_2 \sinh g(x, \bar{q}) \tanh g(x, \bar{q})}{\mathbb{E}_2 \cosh g(x, \bar{q})}\right] \equiv \tilde{q}_1(\bar{q}) \tag{A.29}$$

from Eq. (A.25). Given that $K(x, \bar{q})$ vanishes for $x = \bar{q}$, if the extremum $x = \tilde{q}_1(\bar{q})$ is global in the domain considered, we must have that $K(\tilde{q}_1(\bar{q}), \bar{q}) > 0$ if $x = \tilde{q}_1(\bar{q})$ is a maximum and $K(\tilde{q}_1(\bar{q}), \bar{q}) < 0$ if $x = \tilde{q}_1(\bar{q})$ is a minimum. Therefore, if

$$\partial_{x^2} K(x, \bar{q})|_{x=\tilde{q}_1} = -\frac{\beta'^2 J^2}{4}(P-1)P\tilde{q}_1^{P-2}\left(1 - \frac{\beta'^2 J^2}{2}(P-1)P(\tilde{q}_1(\bar{q}))^{P-2}\mathbb{E}_1\left[\frac{\mathbb{E}_2 \mathrm{sech}^3 g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q})}\right]\right) \tag{A.30}$$

is positive, $K(\tilde{q}_1(\bar{q}), \bar{q})$ is negative and the RS theory becomes unstable. This occurs for

$$1 - \frac{\beta'^2 J^2}{2}(P-1)P(\tilde{q}_1(\bar{q}))^{P-2}\mathbb{E}_1\left[\frac{\mathbb{E}_2 \mathrm{sech}^3\ g(\tilde{q}_1(\bar{q}), \bar{q})}{\mathbb{E}_2 \cosh g(\tilde{q}_1(\bar{q}), \bar{q})}\right] < 0. \tag{A.31}$$

In the limit $\tilde{q}_1(\bar{q}) \to \bar{q}$ this recovers the AT line found in [18]

$$\frac{(P-1)\beta'^2 J^2 P\bar{q}^{P-2}}{2}\mathbb{E}\left[\frac{1}{\cosh^4\sqrt{\beta'^2 J^2 P\bar{q}^{P-1}z/2}}\right] > 1. \tag{A.32}$$

By numerically solving the self-consistency equations (A.19) and (A.25), we can check that the parameters $\bar{q}$ and $\tilde{q}_1(\bar{q})$, are indeed equal at all temperatures see Fig. 4 (mid and right panels), hence the above limit is justified a posteriori.

A.3. P-spin spherical model

In this section we consider the spherical P-spin model, introduced for the first time in [19]. The Hamiltonian of the model is the same as in (A.17), however the N spins σ_i are now real variables, satisfying the so-called 'spherical' constraint $\sum_{i=1}^{N}\sigma_i^2 = N$. As in previous cases, the order parameter is the two-replica overlap q introduced in (2.4). In the thermodynamic limit, at inverse temperature β, the quenched free-energy under the RS assumption (2.6) is given by

$$-2\beta f_{RS}(\bar{q}|\beta) = \frac{\beta^2}{2}(1-\bar{q}^P) + \log(1-\bar{q}) + \frac{\bar{q}}{1-\bar{q}}, \tag{A.33}$$

where $\bar{q}$ fulfills the self-consistency equation

$$\frac{\beta^2}{2}P\bar{q}^{P-1} = \frac{\bar{q}}{(1-\bar{q})^2} \tag{A.34}$$

For later convenience, it is useful to note that the temperature $T^\star$ at which $\bar{q}$ becomes non-zero within the RS theory is found by demanding that (A.34) allows non-zero solutions satisfying

$$\frac{2}{P}T^2 = \bar{q}^{P-2}(1-\bar{q})^2 \tag{A.35}$$

Denoting the RHS with $f(\bar{q})$ and noting that $f(0) = f(1) = 0$, a simple graphical argument shows that a non-zero solution exists when the LHS is smaller than $f(\bar{q})$ evaluated at its maximum point $q = \frac{P-2}{P}$, i.e. for $T < T^\star$ with $T^\star = \sqrt{2(P-2)^{P-2}/P^{P-1}}$ [19,43].

Within the 1RSB assumption (2.12), the quenched free-energy evaluates to

$$\begin{aligned} -2\beta f_{1RSB}(\bar{q}_1,\bar{q}_0|\beta,\theta) =& \frac{\beta^2}{2}\left[1+(\theta-1)\bar{q}_1^P - \theta\bar{q}_0^P\right] \\ &+ \frac{\theta-1}{\theta}\log\left(1-\bar{q}_1\right) + \frac{\bar{q}_0}{1-\bar{q}_1+\theta(\bar{q}_1-\bar{q}_0)} \\ &+ \frac{1}{\theta}\log\left(1-\bar{q}_1+\theta(\bar{q}_1-\bar{q}_0)\right), \end{aligned} \tag{A.36}$$

where $\bar{q}_1$ and $\bar{q}_0$ fulfill the self-consistency equations

$$\frac{\beta^2}{2}P\bar{q}_0^{P-1} = \frac{\bar{q}_0}{(1-\bar{q}_1+\theta(\bar{q}_1-\bar{q}_0))^2} \tag{A.37}$$

$$\frac{\beta^2}{2}P\bar{q}_1^{P-1} = \frac{\beta^2}{2}P\bar{q}_0^{P-1} + \frac{\bar{q}_1-\bar{q}_0}{(1-\bar{q}_1+\theta(\bar{q}_1-\bar{q}_0))^2}. \tag{A.38}$$

From now on, we imply the dependence of f_{RS} and f_{1RSB} on β. We note that for $\theta = 1$, (A.37) becomes equal to (A.34), hence $\bar{q}_0(\theta=1) = \bar{q}$ and we also have $f_{1RSB}(\bar{q}_1,\bar{q}_0|\theta)|_{\theta=1} = f_{RS}(\bar{q})$. The aim is to expand the 1RSB quenched free-energy around $\theta = 1$ to linear orders in $\theta - 1$. To this purpose,

we expand the 1RSB self-consistency equations around $\theta = 1$ to obtain

$$\frac{\beta^2}{2}P\bar{q}_0^{P-1} = \frac{\bar{q}_0}{(1-\bar{q}_0)^2} + (\theta-1)A(\bar{q}_0,\bar{q}_1) \tag{A.39}$$

where $A(\bar{q}_0,\bar{q}_1)$ is defined in (C.17) and

$$\frac{\beta^2}{2}P\bar{q}_1^{P-1} = \frac{\bar{q}_0}{(1-\bar{q}_0)^2} + \frac{\bar{q}_1-\bar{q}_0}{(1-\bar{q}_0)(1-\bar{q}_1)} + (\theta-1)B(\bar{q}_0,\bar{q}_1) \tag{A.40}$$

where $B(\bar{q}_0,\bar{q}_1)$ is as in (C.18). As noted above, if $\theta = 1$, $\bar{q}_0(\theta=1) = \bar{q}$, while $\bar{q}_1(\theta=1)$ fulfills the following equation

$$\frac{\beta^2}{2}P\bar{q}_1^{P-1} = \frac{\bar{q}_1}{(1-\bar{q}_1)} \tag{A.41}$$

We will see below that the RS instability occurs at a temperature $T_c > T^\star$, hence the only solution of (A.37) at T_c is $\bar{q}_0 = 0$. For $\theta = 1$, this corresponds to the paramagnetic solution $\bar{q} = 0$. For $\theta < 1$, the solution $\bar{q}_0 = 0$ remains valid as in the absence of external field all the states must be orthogonal to each other, leading to a vanishing mutual overlap [43]. As the 1RSB theory requires $\bar{q}_1 > \bar{q}_0$, we are interested in the non-zero solution of (A.41), which we can denote with $\tilde{q}_1$ and can be found explicitly from

$$\tilde{q}_1^{P-2}(1-\tilde{q}_1) = \frac{2}{P}T^2 \tag{A.42}$$

Denoting the LHS with $g(\tilde{q}_1)$ and noting that $g(0) = g(1) = 0$, and reasoning as for $T^\star$, a non-zero solution exists when the RHS is smaller than $g(\tilde{q}_1)$ evaluated at its maximum point $\tilde{q}_1^\star = (P-2)/(P-1)$, giving $T < \sqrt{P(P-2)^{P-2}/2(P-1)^{P-1}}$ [43]. Next, we compute the derivative w.r.t. θ of $-2\beta f_{1RSB}$ at $\bar{q}_0 = \tilde{q}$ and $\bar{q}_1 = \tilde{q}_1$:

$$\begin{aligned} K(\tilde{q}_1,\tilde{q}) =& \partial_\theta\left(-2\beta f_{1RSB}(\bar{q}_1,\bar{q}_0|\theta)\right)|_{\theta=1} = \log\left(\frac{1-\tilde{q}_1}{1-\tilde{q}}\right) \\ &+ (\tilde{q}-\tilde{q}_1)\frac{P(1-\tilde{q})(1-2\tilde{q})(1-\tilde{q}_1) - \tilde{q}(1+\tilde{q}) - \tilde{q}_1(1-3\tilde{q})}{P(1-\tilde{q})^3(1-\tilde{q}_1)} \end{aligned} \tag{A.43}$$

Substituting $\bar{q}=0$ and the value of $\bar{q}_1=\bar{q}_1^\star$ this evaluates to

$$K(\bar{q}_1^*,0)=2-\frac{4}{P}-\log(P-1) \tag{A.44}$$

which is always negative for $P>2$, implying

$$f_{1RSB}(\bar{q}_1^*,0|\theta)=f_{RS}(0)+(1-\theta)\left(\frac{K(0,\tilde{q}_1^*)}{2\beta}\right)<f_{RS}(0). \tag{A.45}$$

This shows that at the temperature $T_c=\sqrt{P(P-2)^{P-2}/2(P-1)^{P-1}}$, where a non-zero overlap $\bar{q}_1$ first emerges, the RS theory becomes unstable, as known in the literature [19,43]. It can be easily verified that $T_c>T^\star$ for all $P>2$.

Appendix B. Expanding around $\theta=0$

Although we have so far regarded the limit $\theta\to 1$ (where $\bar{q}_0=\bar{q}$ and $f_{\rm 1RSB}=f_{\rm RS}$) as the physical one, in the opposite limit, $\theta\to 0$, we would *equally* find, for all the models considered above, $f_{\rm 1RSB}=f_{\rm RS}$ (with $\bar{q}_1=\bar{q}$), suggesting that a similar analysis could have been carried for $\theta\to 0$.

In this section we present such analysis for the Hopfield model, the Hebbian networks with multi-node interactions and the spherical P-spin. The same analysis can be carried out for the other spin-glass models considered in Appendix A. Given the strong similarity of the SK and the Ising P-spin models with the Hopfield and the dense associative memory models, respectively, we will not report such analysis here.

B.1. The Hopfield model

From (2.17), one finds

$$\begin{aligned}\lim_{\theta\to 0}\bar{q}_1&=\lim_{\theta\to 0}\mathbb{E}_1\mathbb{E}_2\tanh^2\left(\beta\bar{m}+\beta z^{(1)}\frac{\sqrt{\alpha\bar{q}_0}}{\Delta_2(\theta,\bar{q}_0,\bar{q}_1)}+\beta z^{(2)}\sqrt{\alpha\frac{\bar{q}_1-\bar{q}_0}{\Delta_1(\bar{q}_1)\Delta_2(\theta,\bar{q}_0,\bar{q}_1)}}\right)\\&=\mathbb{E}_1\mathbb{E}_2\tanh^2\left(\beta\bar{m}+\beta z^{(1)}\frac{\sqrt{\alpha\bar{q}_0}}{\Delta_1(\bar{q}_1)}+\beta z^{(2)}\frac{\sqrt{\alpha(\bar{q}_1-\bar{q}_0)}}{\Delta_1(\bar{q}_1)}\right)\\&=\mathbb{E}\tanh^2\left(\beta\bar{m}+\beta\sqrt{\frac{\alpha\bar{q}_1}{(1-\beta(1-\bar{q}_1))^2}}z\right)\end{aligned} \tag{B.1}$$

where we have used that for $\theta=0$, $\Delta_2(0,\bar{q}_0,\bar{q}_1)=\Delta_1(\bar{q}_1)$ and the relation

$$\mathbb{E}_{\lambda,Y}[F(a_1+\lambda a_2+Ya_3)]=\mathbb{E}_Z\left[F\left(a_1+Z\sqrt{a_2^2+a_3^2}\right)\right], \tag{B.2}$$

with F any smooth function, $a_1,\ a_2,\ a_3\in\mathbb{R}$, and λ, Y and Z i.i.d. standard normal random variables. As (B.1) is identical to (2.9), in the limit $\theta\to 0$, $\bar{q}_1$ is equal to the RS order parameter $\bar{q}$. Similarly, one can show that

$$\lim_{\theta\to 0}\bar{m}_1=\bar{m} \tag{B.3}$$

and can easily verify that $f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)|_{\theta=0}=f_{RS}(\bar{m},\bar{q})$. Our purpose is then to prove that for small but finite values of θ the 1RSB expression of the quenched free-energy is smaller than the RS expression, i.e. $f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)<f_{RS}(\bar{m},\bar{q})$, below a critical line in the parameters space (α,β).

To this purpose, we expand the 1RSB quenched free-energy around $\theta=0$ –namely around the replica symmetric expression– to the first order, to write

$$f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)=f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)|_{\theta=0}+\theta\partial_\theta f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)|_{\theta=0}, \tag{B.4}$$

where $f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)|_{\theta=0}=f_{RS}(\bar{m},\bar{q})$. To determine when the RS solution becomes unstable, i.e. $f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)<f_{RS}(\bar{m},\bar{q})$ we inspect the sign of $\partial_\theta f_{1RSB}(\bar{m}_1,\bar{q}_1,\bar{q}_0|\theta)|_{\theta=0}$. To evaluate the latter, we need to expand the self-consistency equations for $\bar{q}_0$, $\bar{q}_1$ and $\bar{m}_1$ around $\theta=0$ to linear orders in θ. Using (B.1) and denoting

$$g_0(\bar{m}_1,\bar{q}_1,\bar{q}_0)=\beta\bar{m}_1+\beta z^{(1)}\frac{\sqrt{\alpha\bar{q}_0}}{\Delta_1(\bar{q}_1)}+\beta z^{(2)}\frac{\sqrt{\alpha(\bar{q}_1-\bar{q}_0)}}{\Delta_1(\bar{q}_1)}, \tag{B.5}$$

we obtain

$$\bar{q}_1=\mathbb{E}_1\mathbb{E}_2\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)+\theta A(\bar{m}_1,\bar{q}_0,\bar{q}_1) \tag{B.6}$$

where $A(\bar{m}_1,\bar{q}_0,\bar{q}_1)$ is a function of $\bar{q}_0$ and $\bar{q}_1$ that will drop out of the calculation, whose expression is provided in (C.4) It follows from (B.1) that to $\mathcal{O}(\theta^0)$, $\bar{q}_1$ is equal to the RS order parameter $\bar{q}$ so we can rewrite (B.6) as

$$\bar{q}_1=\bar{q}+\theta A(\bar{m}_1,\bar{q}_0,\bar{q}). \tag{B.7}$$

Following the same path for $\bar{q}_0$, and using (B.7), we have

$$\bar{q}_0 = \tilde{q}_0(\bar{m}, \bar{q}) + \theta B(\bar{m}_1, \bar{q}_0, \bar{q}) \tag{B.8}$$

where $B(\bar{m}_1, \bar{q}_0, \bar{q})$ is provided in (C.5) and will drop out of the calculation, and we have denoted with $\tilde{q}_0(\bar{m}, \bar{q})$ the solution of

$$\bar{q}_0 = \mathbb{E}_1 \left(\mathbb{E}_2 \tanh g_0(\bar{m}, \bar{q}_0, \bar{q})\right)^2. \tag{B.9}$$

Finally, we can write the magnetization as

$$\bar{m}_1 = \bar{m} + \theta C(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})) \tag{B.10}$$

where $C(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))$ is given in (C.6) and rewrite (B.7) and (B.8) as

$$\bar{q}_1 = \bar{q} + \theta A(\bar{m}, \tilde{q}_0(\bar{m}, \bar{q}), \bar{q}) \tag{B.11}$$

$$\bar{q}_0 = \tilde{q}_0(\bar{m}, \bar{q}) + \theta B(\bar{m}, \tilde{q}_0(\bar{m}, \bar{q}), \bar{q}) \tag{B.12}$$

Using (B.11), (B.10) and (B.12) to evaluate the derivative of $f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta)$ w.r.t. θ and finally setting $\theta = 0$, we obtain:

$$\begin{aligned} &K(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})) := \partial_\theta(-\beta f_{1RSB}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta))|_{\theta=0} \\ &= -\frac{\alpha\beta^2(\bar{q}^2 - \tilde{q}_0(\bar{m}, \bar{q})^2)}{4\Delta_1(\bar{q})^2} + \frac{1}{2}\mathbb{E}_1\mathbb{E}_2 \ln^2 \cosh g_0(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})) - \frac{1}{2}\mathbb{E}_1 \left(\mathbb{E}_2 \ln \cosh g_0(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))\right)^2 \end{aligned} \tag{B.13}$$

Next, we study the sign of (B.13), where $\bar{q}$ and $\tilde{q}_0(\bar{m}, \bar{q})$ are the solutions of the self-consistency equations (2.9) and (B.9), respectively. To this purpose, it is useful to study the behavior of the function $K(\bar{m}, \bar{q}, x)$ for $x \in [0, \bar{q}]$. For $x = \bar{q}$, we have $K(\bar{m}, \bar{q}, \bar{q}) = 0$, regardless of the value assigned to $\bar{m}$, while the extremum of $K(\bar{m}, \bar{q}, x)$ is found from

$$\partial_x K(\bar{m}, \bar{q}, x) = \frac{\beta^2\alpha x}{2\Delta_1(\bar{q})^2}\left[x - \mathbb{E}_1 \left(\mathbb{E}_2 \tanh g_0(\bar{m}, \bar{q}, x)\right)^2\right] = 0$$

as

$$x = \mathbb{E}_1 \left(\mathbb{E}_2 \tanh g_0(\bar{m}, \bar{q}, x)\right)^2 \equiv \tilde{q}_0(\bar{m}, \bar{q}), \tag{B.14}$$

from Eq. (B.9). Given that $K(\bar{m}, \bar{q}, x)$ vanishes for $x = \bar{q}$, if the extremum $x = \tilde{q}_0(\bar{m}, \bar{q})$ is global in the domain considered, we must have that $K(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})) > 0$ if $x = \tilde{q}_0(\bar{m}, \bar{q})$ is a maximum and $K(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})) < 0$ if $x = \tilde{q}_0(\bar{m}, \bar{q})$ is a minimum. Therefore, if

$$\partial_x^2 K(\bar{m}, \bar{q}, x)|_{x=\tilde{q}_0(\bar{m}, \bar{q})} = \frac{\beta^2\alpha}{2\Delta_1(\bar{q})^2}\left\{1 - \frac{\beta^2\alpha}{\Delta_1(\bar{q})^2}\mathbb{E}_1\left\{\mathbb{E}_2\left[\frac{1}{\cosh^2 g_0(\bar{m}, \tilde{q}_0(\bar{m}, \bar{q}), \bar{q})}\right]\right\}^2\right\} \tag{B.15}$$

is negative, $K(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))$ is positive and $f_{1RSB}(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})|\theta) < f_{RS}(\bar{m}, \bar{q})$, hence the RS theory becomes unstable when the expression in the curly brackets in (B.15) becomes negative i.e. for

$$(1 - \beta(1 - \bar{q}))^2 < \beta^2\alpha\mathbb{E}_1\left\{\mathbb{E}_2\left[\frac{1}{\cosh^2 g_0(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))}\right]\right\}^2 \tag{B.16}$$

Interestingly, also in this case, the result found by Coolen in [17] using the de Almeida and Thouless' approach [6], is recovered from the expression above in the limit $\tilde{q}_0(\bar{m}, \bar{q}) \to \bar{q}$. Solving numerically $\bar{q}$ and $\tilde{q}_0(\bar{m}, \bar{q})$ from the self-consistency equations (2.9) and (B.9), respectively, one can verify that these two quantities are indeed identical for any temperature, and the resulting RS instability line coincides with the classical AT line and the critical line given in (2.33), obtained by expanding around $\theta = 1$, see Fig. 5 (left panel). We anticipate that this will remain the case for Hebbian networks with P-node interactions, that we will analyze in the next section (see mid and right panels of Fig. 5). Although we do not report such analysis here, we have checked that this is also the case for the SK model.

B.2. Hebbian networks with multi-node interactions

Here we apply the same analysis to Hebbian networks with multi-node interactions, defined by the Hamiltonian given in Eq. (3.1). Our objective is to prove that the 1RSB quenched free-energy is smaller than its replica symmetric counterpart i.e. $f_{1RSB}(\beta', \alpha, \theta) < f_{RS}(\beta', \alpha)$ above a critical value of the effective parameter $\sqrt{\alpha}\beta'$. To this purpose we expand, to linear orders in θ, the 1RSB quenched free-energy around $\theta = 0$, as shown in (B.4). Since the self-consistency equations also depend on θ, we need to expand them too. Following the same steps as in the Hopfield model, we can write $\bar{m}_1$ as in (B.10), with $C(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))$ as in (C.12), $\bar{q}_1$ as in (B.7), with $A(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))$ given in (C.10), and $\bar{q}_0$ as given in (B.8), where $\tilde{q}_0(\bar{m}, \bar{q})$ is the solution of the self-consistency equation

$$\bar{q}_0 = \mathbb{E}_1 \left(\mathbb{E}_2 \tanh g(\bar{m}, \bar{q}, \bar{q}_0)\right)^2 \tag{B.17}$$

and $B(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q}))$ is given in (C.11). With the above expressions in hand, we can now calculate the derivative of $f_{1\text{RSB}}$ w.r.t. θ when $\theta = 0$, as needed in (B.4)

$$K(\bar{m}, \bar{q}, \tilde{q}_0(\bar{m}, \bar{q})) := \partial_\theta(-\beta' f_{1\text{RSB}}(\bar{m}_1, \bar{q}_1, \bar{q}_0|\theta))|_{\theta=0}$$

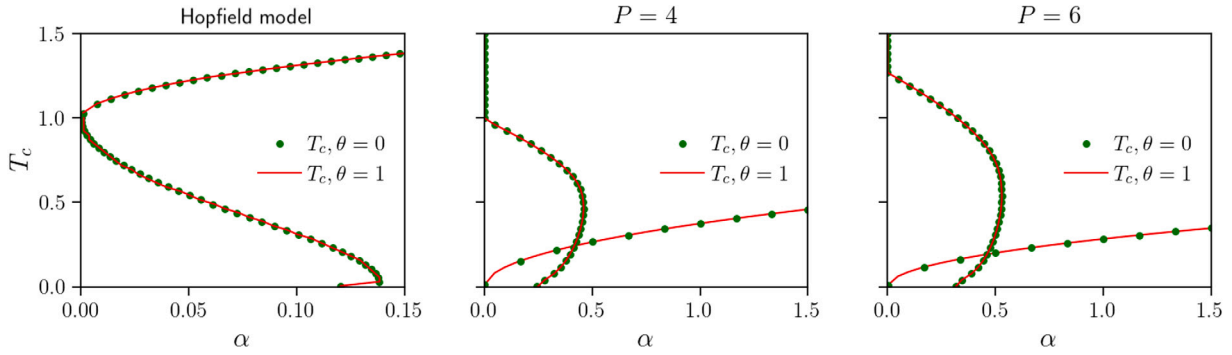

Fig. 5. RS instability lines for the Hopfield model (left) and Hebbian networks with P-node interactions, with $P = 4$ (mid) and $P = 6$ (right), obtained via our method for $\theta \to 0$ and $\theta \to 1$. In the all the case the lines obtained for $\theta \to 0$ and $\theta \to 1$ are indistinguishable.

$$= \frac{-\alpha\beta'^2(P-1)}{4}(\bar{q}^P - (\tilde{q}_0(\bar{m},\bar{q}))^P) + \frac{1}{2}\mathbb{E}_1\mathbb{E}_2 \ln^2\cosh g(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q})) - \frac{1}{2}\mathbb{E}_1\left(\mathbb{E}_2 \ln\cosh g(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q}))\right)^2 \tag{B.18}$$

Again, we have that $K(\bar{m},\bar{q},\bar{q}) = 0$, regardless of the value assigned to $\bar{m}$ (this follows from the fact that for $\theta = 0$, $\bar{q}$ is an extremum of the free-energy). Next, we inspect the sign of $K(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q}))$. To this purpose, we study $K(\bar{m},\bar{q},x)$ for $x \in [0,\bar{q}]$ and locate its extrema, which are found from

$$\partial_x K(\bar{m},\bar{q},x) = \frac{\beta'^2\alpha P(P-1)}{4}x^{P-2}\left[x - \mathbb{E}_1\left(\mathbb{E}_2 \tanh g(\bar{m},\bar{q},x)\right)^2\right] = 0 \tag{B.19}$$

as

$$x = \mathbb{E}_1\left(\mathbb{E}_2 \tanh g(\bar{m},\bar{q},x)\right)^2 \equiv \tilde{q}_0(\bar{m},\bar{q}) \tag{B.20}$$

where the last equality follows from (B.17). Under the assumption that the extremum $x = \tilde{q}_0(\bar{m},\bar{q})$ is global in the domain considered and reasoning as in the Hopfield case, we have that $K(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q})) > 0$ if $x = \tilde{q}_0(\bar{m},\bar{q})$ is a maximum and $K(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q})) < 0$ if it is a minimum. In particular, if

$$\partial_x^2 K(\bar{m},\bar{q},x)|_{x=\tilde{q}_0(\bar{m},\bar{q})} = -\frac{\beta'^2\alpha P(P-1)}{4}\tilde{q}_0(\bar{m},\bar{q})^{P-2} \cdot\left\{1 - \frac{\beta'^2\alpha P(P-1)(\tilde{q}_0(\bar{m},\bar{q}))^{P-2}}{2}\mathbb{E}_1\left[\mathbb{E}_2\frac{1}{\cosh^2 g(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q}))}\right]^2\right\} \tag{B.21}$$

is negative, $K(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q})) > 0$ and $f_{1RSB} < f_{RS}$. This happens when the expression in the curly brackets of the equation above is negative, i.e. when the parameter $\alpha\beta'^2$ satisfies the inequality

$$\frac{\alpha\beta'^2 P(P-1)(\tilde{q}_0(\bar{m},\bar{q}))^{P-2}}{2}\mathbb{E}_1\left\{\mathbb{E}_2\left[\frac{1}{\cosh^2 g(\bar{m},\bar{q},\tilde{q}_0(\bar{m},\bar{q}))}\right]\right\}^2 > 1. \tag{B.22}$$

The resulting critical line is found to be identical to the critical line (3.11) obtained from the expansion around $\theta = 1$ (see Fig. 5, mid and right panels).

B.3. Spherical P-spin

Expanding for small θ the self-consistency equations (A.37), (A.38) to linear orders, we get

$$\frac{\beta^2}{2}P\bar{q}_1^{P-1} = \frac{\beta^2}{2}P\bar{q}_0^{P-1} + \frac{\bar{q}_1 - \bar{q}_0}{(1-\bar{q}_1)^2} + \theta A(\bar{q}_0,\bar{q}_1) \tag{B.23}$$

$$\frac{\beta^2}{2}P\bar{q}_0^{P-1} = \frac{\bar{q}_0}{(1-\bar{q}_1)^2} + \theta B(\bar{q}_0,\bar{q}_1) \tag{B.24}$$

where the expression for $A(\bar{q}_0,\bar{q}_1)$ and $B(\bar{q}_0,\bar{q}_1)$ are provided in (C.19) and (C.20), respectively. If $\theta = 0$, summing the two equations gives

$$\frac{\beta^2}{2}P\bar{q}_1^{P-1} = \frac{\bar{q}_1}{(1-\bar{q}_1)^2} \tag{B.25}$$

showing that, to orders $\mathcal{O}(\theta^0)$, $\bar{q}_1 = \bar{q}$, while $\bar{q}_0$ fulfills the following self-consistency equation

$$\frac{\beta^2}{2}P\bar{q}_0^{P-1} = \frac{\bar{q}_0}{(1-\bar{q}_1)^2} \tag{B.26}$$

whose solution is denoted with $\tilde{q}_0(\bar{q})$. The latter equation is solved by $\tilde{q}_0(\bar{q}) = 0$ (which corresponds to the paramagnetic solution and remains valid when ergodicity is broken, as explained earlier). Similarly, $\bar{q}_1 = 0$ is always a solution of (B.25), however the 1RSB scenario requires $\bar{q}_1 > 0$. As explained in the previous section, such non-zero solution appears at $T \leq T^\star = \sqrt{2(P-2)^{P-2}/P^{P-1}}$, which is below T_c for any $P > 2$, hence we can immediately conclude that the instability of the RS theory occurs at the larger temperature T_c, without further comparing the free-energies 1RSB and RS for θ close to zero.

Appendix C. Contributions to sub-leading orders

In this appendix we provide expressions for all the functions that we left unspecified in the main text, as they did not contribute to leading orders, including the functions $A(\bar{q}_0, \bar{q}_1)$ and $B(\bar{q}_0, \bar{q}_1)$ for all the models considered.

- For the Hopfield model, in the expansion around $\theta = 1$ the subleading contributions to (2.22) and (2.24) are,

$$A(\bar{m}, \bar{q}_0, \bar{q}_1) = 2\mathbb{E}_1 \left[\tanh\left(\beta\bar{m} + \frac{\beta z^{(1)} \sqrt{\alpha \bar{q}_0}}{(1-\beta(1-\bar{q}_0))} \right) \cdot \frac{\mathbb{E}_2 \log\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \sinh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)(1 - \tanh g_1(\bar{m}, \bar{q}_0, \bar{q}_1))}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right], \tag{C.1}$$

$$\begin{aligned} B(\bar{m}, \bar{q}_0, \bar{q}_1) = {} & \frac{\beta^3 \alpha (\bar{q}_1 - \bar{q}_0)}{(1-\beta(1-\bar{q}_0))^2} \left(\frac{\bar{q}_0}{(1-\beta(1-\bar{q}_0))} + \frac{(\bar{q}_1 - \bar{q}_0)}{(1-\beta(1-\bar{q}_1))} \right) \\ & \cdot \mathbb{E}_1 \left\{ 1 + \frac{\mathbb{E}_2 \tanh^2 g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} + \frac{\mathbb{E}_2 \log\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)(1 - \tanh^2 g_1(\bar{m}, \bar{q}_0, \bar{q}_1))}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right\} \\ & - \frac{\beta^3 \alpha (\bar{q}_1 - \bar{q}_0)^2}{(1-\beta(1-\bar{q}))^3} \mathbb{E}_1 \left\{ \tanh^2 g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \left(1 - \frac{\mathbb{E}_2 \sinh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh^2 g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right) \right. \\ & + \left(2 - 3\tanh^3 \left(\beta\bar{m} + \frac{\beta z^{(1)} \sqrt{\alpha \bar{q}_0}}{(1-\beta(1-\bar{q}_0))} \right) \right) \frac{\mathbb{E}_2 \sinh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh^2 g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \\ & \left. + \tanh\left(\beta\bar{m} + \frac{\beta z^{(1)} \sqrt{\alpha \bar{q}_0}}{(1-\beta(1-\bar{q}_0))} \right) \frac{\mathbb{E}_2 \log\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right\} \\ & \frac{\beta^3 \alpha (\bar{q}_1 - \bar{q}_0)^2}{2(1-\beta(1-\bar{q}_0))^2 (1-\beta(1-\bar{q}_1))} \mathbb{E}_1 \left\{ \tanh\left(\beta\bar{m} + \frac{\beta z^{(1)} \sqrt{\alpha \bar{q}_0}}{(1-\beta(1-\bar{q}_0))} \right) \right. \\ & \left. \cdot \frac{\mathbb{E}_2 \sinh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right\} \\ & - \mathbb{E}_1 \left\{ \frac{\mathbb{E}_2 \sinh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \log\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{(\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1))^2} \right\}, \end{aligned} \tag{C.2}$$

and

$$\begin{aligned} C(\bar{m}, \bar{q}, \bar{q}_1) = {} & \left(-3\beta \frac{\sqrt{\alpha \bar{q}}}{\Delta_1(\bar{q})} - \beta \sqrt{\frac{\alpha(\bar{q}_1 - \bar{q})}{\Delta_1(\bar{q}) \Delta_1(\bar{q}_1)}} \right) \mathbb{E} \tanh\left(\beta\bar{m} + \beta \sqrt{\frac{\alpha \bar{q}}{(1-\beta(1-\bar{q}))^2}} z \right) \\ & + 2\beta \mathbb{E} \tanh^3 \left(\beta\bar{m} + \beta \sqrt{\frac{\alpha \bar{q}}{(1-\beta(1-\bar{q}))^2}} z \right) \frac{\sqrt{\alpha \bar{q}}}{\Delta_1(\bar{q})} \\ & - \beta \frac{\sqrt{\alpha \bar{q}}}{\Delta_1(\bar{q})} \mathbb{E}_1 \left[\sinh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \frac{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \log\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right] \\ & + \left(\beta \frac{\sqrt{\alpha \bar{q}}}{\Delta_1(\bar{q})} + \beta \sqrt{\frac{\alpha(\bar{q}_1 - \bar{q})}{\Delta_1(\bar{q}) \Delta_1(\bar{q}_1)}} \right) \left(\mathbb{E}_1 \left[\frac{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \log\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right] \right. \\ & \left. + \mathbb{E}_1 \left[\frac{2\cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1) \tanh^2 g_1(\bar{m}, \bar{q}_0, \bar{q}_1)}{\mathbb{E}_2 \cosh g_1(\bar{m}, \bar{q}_0, \bar{q}_1)} \right] \right) \end{aligned} \tag{C.3}$$

respectively, where $g_1(\bar{m}, \bar{q}_0, \bar{q}_1)$ is defined as in (2.18).
For the expansion around $\theta = 0$ the subleading contributions to (B.6) and (B.8) are

$$\begin{aligned} A(\bar{m}, \bar{q}_0, \bar{q}_1) = {} & \mathbb{E}_1 \left[\mathbb{E}_2 \tanh^2 g_0(\bar{m}, \bar{q}_1, \bar{q}_0) \ln\cosh g_0(\bar{m}, \bar{q}_1, \bar{q}_0) \right] \\ & - \mathbb{E}_1 \left[\mathbb{E}_2 \tanh^2 \mathbb{E}_2 \ln\cosh g_0(\bar{m}, \bar{q}_1, \bar{q}_0) \right] \\ & + \frac{\beta^3 \alpha \bar{q}_1 (\bar{q}_1 - \bar{q}_0)}{\Delta_1(\bar{q}_1)^3} \left[1 - \mathbb{E}_1 \left(\mathbb{E}_2 \tanh^2 g_0(\bar{m}, \bar{q}_1, \bar{q}_0) \right) \right] \end{aligned}$$

$$+\frac{3\beta^3\alpha(\bar{q}_1^2-\bar{q}_0^2)}{\Delta_1(\bar{q}_1)^3}\mathbb{E}_1\left[\mathbb{E}_2\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_2)\left(1-\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right)\right] \tag{C.4}$$

$$\begin{aligned}B(\bar{m},\bar{q}_0,\bar{q}_1) = 2\Big\{&\mathbb{E}_1\left[\mathbb{E}_2\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\log\cosh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right]\\
&-\mathbb{E}_1\left[\mathbb{E}_2\log\cosh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\left(\mathbb{E}_2\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right)^2\right]\\
&+\frac{2\beta^2\alpha\bar{q}_0(\bar{q}_1-\bar{q}_0)}{\Delta_1(\bar{q}_1)^3}\mathbb{E}_1\left[\mathbb{E}_2\left(1-\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right)\right]\\
&-\frac{4\beta^3\alpha\bar{q}_0(\bar{q}_1-\bar{q}_0)}{\Delta_1(\bar{q}_1)^3}\mathbb{E}_1\left[\mathbb{E}_2\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)(1-\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0))\mathbb{E}_2\tanh^2 g_0(\bar{q}_1,\bar{q}_0)\right]\\
&+\frac{2\beta^3\alpha\bar{q}_0(\bar{q}_1-\bar{q}_0)}{\Delta_1(\bar{q}_1)^3}\mathbb{E}_1\left[\mathbb{E}_2\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\left(1-\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right)\right]\\
&+\frac{\beta^2\alpha(\bar{q}_1-\bar{q}_0)^2}{\Delta_1(\bar{q}_1)^3}\mathbb{E}_1\left[\mathbb{E}_2\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)(1-\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0))\right]\Big\}\end{aligned} \tag{C.5}$$

and

$$\begin{aligned}C(\bar{m},\bar{q},\bar{q}_0) =&\frac{\alpha\beta^3\bar{q}_1(\bar{q}_1-\bar{q}_0)}{(1-\beta(1-\bar{q}_1))^3}\mathbb{E}\left[\left(1-\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right)\log\cosh g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right.\\
&\left.+\tanh^2 g_0(\bar{m},\bar{q}_1,\bar{q}_0)-2\tanh g_0(\bar{m},\bar{q}_1,\bar{q}_0)+2\tanh^3 g_0(\bar{m},\bar{q}_1,\bar{q}_0)\right]\end{aligned} \tag{C.6}$$

where $g_0(\bar{m},\bar{q}_0,\bar{q}_1)$ is defined in (B.5).

- For Hebbian networks with P-node interactions, the subleading contributions to the overlaps $\bar{q}_0$ and $\bar{q}_1$ in the expansion around $\theta=1$ are given by

$$\begin{aligned}A(\bar{m},\bar{q}_0,\bar{q}_1) =&2\mathbb{E}_1\left\{\frac{\mathbb{E}_2\ln\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)}{\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)}\right\}-\\
&2\mathbb{E}_1\left\{\frac{\mathbb{E}_2\ln\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)}{\left(\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right)^2}\right\},\end{aligned} \tag{C.7}$$

$$\begin{aligned}B(\bar{m},\bar{q}_0,\bar{q}_1) =&2\mathbb{E}_1\left\{\frac{\mathbb{E}_2\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\log\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)}{\left(\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right)^2}\right\}\\
&-2\mathbb{E}_1\left\{\frac{\left(\mathbb{E}_2\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\right)^2\mathbb{E}_2\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\log\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)}{\left(\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right)^3}\right\},\end{aligned} \tag{C.8}$$

$$\begin{aligned}C(\bar{m},\bar{q}_0,\bar{q}_1) =&\mathbb{E}_1\left[\frac{\mathbb{E}_2\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\log\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)}{\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)}\right]-\\
&\mathbb{E}_1\left[\frac{\mathbb{E}_2\sinh g(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\log\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)}{\left(\mathbb{E}_2\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right)^2}\right]\end{aligned} \tag{C.9}$$

respectively, whereas, for the expansion around $\theta=0$ they evaluate to

$$\begin{aligned}A(\bar{m},\bar{q}_0,\bar{q}_1) =&\mathbb{E}_1\mathbb{E}_2\ln\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\tanh^2 g(\bar{m},\bar{q}_1,\bar{q}_0)\\
&-\mathbb{E}_1\left(\mathbb{E}_2\ln\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\tanh^2 g(\bar{m},\bar{q}_1,\bar{q}_0)\right)\end{aligned} \tag{C.10}$$

$$\begin{aligned}B(\bar{m},\bar{q}_0,\bar{q}_1) =&\left\{2\mathbb{E}_1\left[\mathbb{E}_2\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\ln\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\right]\right.\\
&\left.-2\mathbb{E}_1\left[\left(\mathbb{E}_2\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\right)^2\mathbb{E}_2\ln\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right]\right\},\end{aligned} \tag{C.11}$$

$$\begin{aligned}C(\bar{m},\bar{q}_0,\bar{q}_1) =&\mathbb{E}_1\left[\mathbb{E}_2\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\log\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right]\\
&-\mathbb{E}_1\left[\mathbb{E}_2\tanh g(\bar{m},\bar{q}_1,\bar{q}_0)\mathbb{E}_2\log\cosh g(\bar{m},\bar{q}_1,\bar{q}_0)\right]\end{aligned} \tag{C.12}$$

where $g(\bar{m},\bar{q}_1,\bar{q}_0)$ is defined in (3.5).

- For the Sherrington-Kirkpatrick model the subleading contributions to the overlaps $\bar{q}_0$ and $\bar{q}_1$ in the expansion around $\theta=1$ are given by

$$\begin{aligned}A(\bar{q}_0,\bar{q}_1) =\mathbb{E}_1\Bigg\{&\frac{\mathbb{E}_2\log\cosh g(\bar{q}_0,\bar{q}_1)\sinh g(\bar{q}_0,\bar{q}_1)\tanh g(\bar{q}_0,\bar{q}_1)\mathbb{E}_2\cosh g(\bar{q}_0,\bar{q}_1)}{(\mathbb{E}_2\cosh g)^2}\\
&-\frac{\mathbb{E}_2\sinh g(\bar{q}_0,\bar{q}_1)\tanh g(\bar{q}_0,\bar{q}_1)\mathbb{E}_2\cosh g(\bar{q}_0,\bar{q}_1)\log\cosh g(\bar{q}_0,\bar{q}_1)}{(\mathbb{E}_2\cosh g(\bar{q}_0,\bar{q}_1))^2}\Bigg\}\end{aligned} \tag{C.13}$$

$$B(\bar{q}_0,\bar{q}_1) = \mathbb{E}_1 \left\{ \frac{\mathbb{E}_2 \log\cosh g(\bar{q}_0,\bar{q}_1) \sinh g(\bar{q}_0,\bar{q}_1) \mathbb{E}_2 \cosh g(\bar{q}_0,\bar{q}_1)}{(\mathbb{E}_2 \cosh g(\bar{q}_0,\bar{q}_1))^2} - \frac{\mathbb{E}_2 \sinh g(\bar{q}_0,\bar{q}_1) \mathbb{E}_2 \cosh g(\bar{q}_0,\bar{q}_1) \log\cosh g(\bar{q}_0,\bar{q}_1)}{(\mathbb{E}_2 \cosh g(\bar{q}_0,\bar{q}_1))^2} \right\}. \tag{C.14}$$

where $g(\bar{q}_1,\bar{q}_0)$ is defined in (A.5).

- For the Ising P-spin model, these terms, in the expansion around $\theta = 1$, evaluate to

$$A(\bar{q}_0,\bar{q}_1) = \mathbb{E}_1 \left\{ \frac{\mathbb{E}_2 \log\cosh g(\bar{q}_1,\bar{q}_0) \sinh g(\bar{q}_1,\bar{q}_0) \tanh g(\bar{q}_1,\bar{q}_0) \mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0)}{(\mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0))^2} - \frac{\mathbb{E}_2 \sinh g(\bar{q}_1,\bar{q}_0) \tanh g(\bar{q}_1,\bar{q}_0) \mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0) \log\cosh g(\bar{q}_1,\bar{q}_0)}{(\mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0))^2} \right\} \tag{C.15}$$

$$B(\bar{q}_0,\bar{q}_1) = \mathbb{E}_1 \left\{ \frac{\mathbb{E}_2 \log\cosh g(\bar{q}_1,\bar{q}_0) \sinh g(\bar{q}_1,\bar{q}_0) \mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0)}{(\mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0))^2} - \frac{\mathbb{E}_2 \sinh g(\bar{q}_1,\bar{q}_0) \mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0) \log\cosh g(\bar{q}_1,\bar{q}_0)}{(\mathbb{E}_2 \cosh g(\bar{q}_1,\bar{q}_0))^2} \right\}. \tag{C.16}$$

where $g(\bar{q}_1,\bar{q}_0)$ is defined in (A.21).

- Finally, for the spherical P-spin model, the contributions to linear orders in the expansion around $\theta = 1$ (see (A.39) and (A.40)) are

$$A(\bar{q}_0,\bar{q}_1) = \frac{2\bar{q}_0(\bar{q}_0-\bar{q}_1)}{(1-\bar{q}_0)^3} \tag{C.17}$$

$$B(\bar{q}_0,\bar{q}_1) = \frac{(\bar{q}_1-\bar{q}_0)^2}{(1-\bar{q}_0)(1-\bar{q}_1)} + \frac{2\bar{q}_0(\bar{q}_0-\bar{q}_1)}{(1-\bar{q}_0)^3}. \tag{C.18}$$

while the contributions to linear orders in the expansion around $\theta = 0$ (see (B.23) and (B.24)) are:

$$A(\bar{q}_0,\bar{q}_1) = \frac{(\bar{q}_1^2-\bar{q}_0^2)}{(1-\bar{q}_1)^3} \tag{C.19}$$

$$B(\bar{q}_0,\bar{q}_1) = \frac{2\bar{q}_0(\bar{q}_1-\bar{q}_0)}{(1-\bar{q}_1)^3} \tag{C.20}$$

References

[1] L. Albanese, F. Alemanno, A. Alessandrelli, A. Barra, Replica symmetry breaking in dense hebbian neural networks, J. Stat. Phys. 189 (2) (2022) 1–41.
[2] C. Baldassi, C. Borgs, J.T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, R. Zecchina, Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, Proc. Natl. Acad. Sci. 113 (48) (2016) E7655–E7662.
[3] C. Baldassi, C. Lauditi, E.M. Malatesta, G. Perugini, R. Zecchina, Unveiling the structure of wide flat minima in neural networks, Phys. Rev. Lett. 127 (27) (2021) 278301.
[4] Y. Zhao, J. Qiu, M. Xie, H. Huang, Equivalence between belief propagation instability and transition to replica symmetry breaking in perceptron learning systems, Phys. Rev. Res. 4 (2) (2022) 023023.
[5] G.S. Hartnett, E. Parker, E. Geist, Replica symmetry breaking in bipartite spin glasses and neural networks, Phys. Rev. E 98 (2) (2018) 022116.
[6] J.R. de Almeida, D.J. Thouless, Stability of the Sherrington–Kirkpatrick solution of a spin glass model, J. Phys. A: Math. Gen. 11 (5) (1978) 983.
[7] M. Talagrand, et al., Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models, Vol. 46, Springer Science & Business Media, 2003.
[8] X. Bardina, D. Márquez-Carreras, C. Rovira, S. Tindel, The p-spin interaction model with external field, Potential Anal. 21 (2004) 311–362.
[9] F. Guerra, The replica symmetric region in the Sherrington–Kirkpatrick mean field spin glass model. the almeida-thouless line, 2006, arXiv preprint cond-mat/0604674.
[10] W.-K. Chen, On the Almeida–Thouless transition line in the Sherrington–Kirkpatrick model with centered Gaussian external field, Electron. Commun. Probab. 26 (2021) 1–9.
[11] J. Höller, N. Read, One-step replica-symmetry-breaking phase below the de Almeida–Thouless line in low-dimensional spin glasses, Phys. Rev. E 101 (4) (2020) 042114.
[12] C. Manai, S. Warzel, The de Almeida–Thouless line in hierarchical quantum spin glasses, J. Stat. Phys. 186 (1) (2022) 1–32.
[13] P. Charbonneau, Y. Hu, A. Raju, J.P. Sethna, S. Yaida, Morphology of renormalization-group flow for the de Almeida–Thouless–Gardner universality class, Phys. Rev. E 99 (2) (2019) 022132.
[14] T. Temesvári, I. Kondor, Field theory for the Almeida–Thouless transition, 2022, arXiv preprint arXiv:2212.01654.
[15] F.L. Toninelli, About the Almeida–Thouless transition line in the Sherrington–Kirkpatrick mean-field spin glass model, Europhys. Lett. 60 (5) (2002) 764.
[16] F. Guerra, Broken replica symmetry bounds in the mean field spin glass model, Comm. Math. Phys. 233 (2003) 1–12.
[17] A. Coolen, Statistical mechanics of recurrent neural networks i—statics, in: Handbook of Biological Physics, Vol. 4, Elsevier, 2001, pp. 553–618.
[18] E. Gardner, Spin glasses with p-spin interactions, Nuclear Phys. B 257 (1985) 747–765.
[19] A. Crisanti, H.J. Sommers, The spherical p-spin interaction spin glass model: the statics, Z. Phys. B 87 (1992) 341–354.
[20] D.J. Amit, Modeling Brain Function: The World of Attractor Neural Networks, Cambridge University Press, 1989.
[21] A.C.C. Coolen, R. Kühn, P. Sollich, Theory of Neural Information Processing Systems, OUP Oxford, 2005.
[22] A. Crisanti, D.J. Amit, H. Gutfreund, Saturation level of the hopfield model for neural network, Europhys. Lett. (EPL) 2 (1986) 337–341.
[23] H. Steffan, R. Kühn, Replica symmetry breaking in attractor neural network models, Z. Phys. B 95 (1994).
[24] E. Agliari, L. Albanese, A. Barra, G. Ottaviani, Replica symmetry breaking in neural networks: A few steps toward rigorous results, J. Phys. A 53 (2020).
[25] D. Krotov, J.J. Hopfield, Dense associative memory for pattern recognition, in: Advances in Neural Information Processing Systems, 2016, pp. 1180–1188.
[26] E. Agliari, F. Alemanno, A. Barra, A. Fachechi, Generalized Guerra's interpolation schemes for dense associative neural networks, Neural Netw. 128 (2020) 254–267.

[27] E. Agliari, F. Alemanno, A. Barra, M. Centonze, A. Fachechi, Neural networks with a redundant representation: Detecting the undetectable, Phys. Rev. Lett. 124 (2020) 28301.
[28] E. Gardner, Multiconnected neural network models, J. Phys. A: Gen. Phys. 20 (1987).
[29] A. Barra, G. Genovese, F. Guerra, D. Tantari, How glassy are neural networks? J. Stat. Mech. Theory Exp. 2012 (07) (2012) P07009.
[30] A. Annibale, G. Gualdi, A. Cavagna, Coexistence of supersymmetric and supersymmetry-breaking states in spherical spin-glasses, J. Phys. A: Math. Gen. 37 (47) (2004) 11311, 37 11311.
[31] A. Crisanti, L. Leuzzi, Spherical $2+p$ spin-glass model: An exactly solvable model for glass to spin-glass transition, Phys. Rev. Lett. 93 (2004) 217203.
[32] G. Folena, S. Franz, F. Ricci-Tersenghi, Rethinking mean-field glassy dynamics and its relation with the energy landscape: The surprising case of the spherical mixed p-spin model, Phys. Rev. X 10 (2020) 031045.
[33] A. Crisanti, L. Leuzzi, Spherical $2+p$ spin-glass model: An analytically solvable model with a glass-to-glass transition, Phys. Rev. B 73 (2006) 014412.
[34] T. Tonolo, J. Niedda, G. Gradenigo, Marginal stability in the spherical spin glass: on the competition between disorder and non-linearity, 2023, in preparation.
[35] F. Antenucci, A. Crisanti, L. Leuzzi, Complex spherical 2 + 4 spin glass: A model for nonlinear optics in random media, Phys. Rev. A 91 (2015) 053816.
[36] F. Antenucci, C. Conti, A. Crisanti, L. Leuzzi, General phase diagram of multimodal ordered and disordered lasers in closed and open cavities, Phys. Rev. Lett. 114 (2015) 043901.
[37] F. Antenucci, M. Ibáñez Berganza, L. Leuzzi, Statistical physics of nonlinear wave interaction, Phys. Rev. B 92 (2015) 014204.
[38] F. Antenucci, A. Crisanti, M. Ibáñez Berganza, A. Marruzzo, L. Leuzzi, Statistical mechanics models for multimode lasers and random lasers, Phil. Mag. 96 (7–9) (2016) 704–731.
[39] F. Antenucci, G. Lerario, B.S. Fernandéz, L. De Marco, M. De Giorgi, D. Ballarini, D. Sanvitto, L. Leuzzi, Demonstration of self-starting nonlinear mode locking in random lasers, Phys. Rev. Lett. 126 (2021) 173901.
[40] D.J. Thouless, Spin-glass on a bethe lattice, Phys. Rev. Lett. 56 (1986) 1082–1085.
[41] D. Sherrington, S. Kirkpatrick, Solvable model of a spin-glass, Phys. Rev. Lett. 35 (26) (1975) 1792.
[42] H. Nishimori, Statistical Physics of Spin Glasses and Information Processing: An Introduction. Number 111, Clarendon Press, 2001.
[43] T. Castellani, A. Cavagna, Spin-glass theory for pedestrians, J. Stat. Mech. Theory Exp. 2005 (05) (2005) P05012.