Equivalence between belief propagation instability and transition to replica symmetry breaking in perceptron learning systems

Yang Zhao, Junbin Qiu, Mingshan Xie, and Haiping Huang*
PMI Lab, School of Physics, Sun Yat-sen University, Guangzhou 510275, People's Republic of China

(Received 15 December 2021; accepted 22 March 2022; published 8 April 2022)

The binary perceptron is a fundamental model of supervised learning for nonconvex optimization, which is a root of the popular deep learning. The binary perceptron is able to achieve a classification of random high-dimensional data based on the marginal probabilities of binary synapses. The relationship between the belief propagation instability and the equilibrium analysis of the model remains elusive. Here, we establish the relationship by showing that the instability condition around the belief propagation fixed point is identical to the instability for breaking the replica symmetric saddle-point solution of the free-energy function. Therefore our analysis will hopefully provide insight towards other learning systems in bridging the gap between nonconvex learning dynamics and statistical mechanics properties of more complex neural networks.

DOI: 10.1103/PhysRevResearch.4.023023

I. INTRODUCTION

Theoretical studies of neural networks have become increasingly important in recent years [1–3], as deep neural networks are widely used in various domains of both scientific and industrial communities. One of the most powerful theoretical tools is the replica method, which is able to derive equilibrium properties of neural networks (systems of interacting neurons or synapses), such as the phase diagram [4–7], storage capacity [8–11], and even large-deviation behavior of learning algorithms [12–14]. Intuitively, the replica method introduces n (an integer) copies of the original system. Within each copy, there exist strong interactions among constituent elements (e.g., synaptic or neural states), and these interactions make the model intractable without any approximation in most cases. However, the elements would become decoupled from each other as an overlap (of states) matrix is introduced, which allows a hierarchical level of approximation depending on the stability analysis of the saddle points of the free-energy action. A seminal approximation, namely, replica symmetry breaking, was introduced by Parisi in the 1980s [15,16].

Overall, the replica method, despite its nonintuitive physics, could lead to exact results in some models. One drawback of this method is that it could not be used to design any efficient algorithms in neural networks. Instead, the cavity method is constructed in a physically intuitive way, i.e., a statistical mechanics model of learning can be mapped onto a graphical model, where interactions are represented by factor nodes and synapses are represented by variable nodes (such as the graphical model representation in unsupervised learning [17]). Through virtually deleting these two kinds of nodes, a cavity probability could be defined. Using the treelike structures of the factor graph, or the weakly-interacting-element assumption, an iterative equation for these cavity probabilities can be derived, which leads to self-consistent evaluations of thermodynamic quantities, such as ground-state energy, free energy, and entropy [18–20]. Most interestingly, this iterative equation is exactly the same as the belief propagation developed independently in computer science [21]. Belief propagation could also be derived for learning problems with discrete synapses [22]. An open question is whether the replica symmetry breaking transition corresponds to the belief propagation instability in the learning of neural networks.

Here, we provide a proof of this fundamental equivalence in the seminal model of binary perceptron learning, in which learning is achieved by adjusting discrete synapses (actually, the synaptic state takes ± 1). This model was first studied by Gardner and Derrida [10,23]. A follow-up calculation showed that the storage capacity of this model is given by $P_c \simeq 0.833N$ [11], where N is the number of neurons and P_c is the critical number of random patterns being correctly classified. The binary perceptron belongs to the NP-hard class in the worst-case complexity. The typical weight configuration is quite hard to find by any algorithms based on local flips (e.g., Monte Carlo dynamics) [24–27]. The first efficient algorithm was inspired by the cavity method, reaching an algorithmic threshold $P_{\rm alg} \simeq 0.72N$ [22]. It was then proved by defining a distance-dependent potential that the entire solution space is composed of single valleys of vanishing entropy [28]. This picture was further shown to be mathematically rigorous in some perceptron learning problems [29]. However, the region of the solution space accessed by practical algorithms does not belong to the equilibrium hard-to-reach isolated parts, but to the subdominant dense parts [12,13]. These dense parts are

*huanghp7@mail.sysu.edu.cn

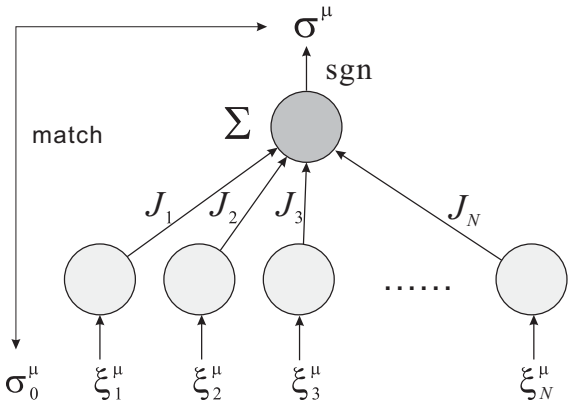

FIG. 1. Sketch of a binary perceptron. The binary perceptron is composed of N input units (light gray circles) connected to one output unit (dark gray circle). Each input unit receives one pixel of the input pattern. The output unit computes a weighted sum (indicated by Σ), which is passed through a nonlinear sign function (indicated by "sgn") to carry out a binary classification. The task of the binary perceptron is to find a set of weights that can match all the actual classification ($\{\sigma^\mu\}$) of given patterns with their labels ($\{\sigma_0^\mu\}$).

further shown to have good generalization properties [14,30], providing a new paradigm to understand deep learning.

Therefore studying the mathematical foundation of the binary perceptron problem is fundamentally important to our understanding of neural networks. To our best knowledge, studies on the relationship between the replica symmetry breaking transition and the belief propagation instability along this line are rare. In this paper, we show how the belief propagation instability is connected to the instability of the replica symmetric (RS) solution (or replicon mode) of the model.

II. BINARY PERCEPTRON

The binary perceptron is a single-layer neural network that learns a random input-output mapping by discrete synapses (see Fig. 1). We assume that there are P uncorrelated input-output associations, where the μth one consists of an N-dimensional pattern $\boldsymbol{\xi}^\mu$ and a corresponding label σ_0^μ, where ξ_i^μ and σ_0^μ take ± 1 with equal probability. Given a configuration of synaptic weights $\{J_i\}_{i=1}^N$ (each entry takes $+1$ or -1), the binary perceptron gives the output $\sigma^\mu = \mathrm{sgn}\,(\sum_{i=1}^N J_i\xi_i^\mu)$ for the input pattern $\boldsymbol{\xi}^\mu$. If $\sigma^\mu = \sigma_0^\mu$, we say that the synaptic weight vector $\mathbf{J}$ has recognized the μth pattern. The binary perceptron is able to store an extensive number of random patterns. Therefore we define a loading rate $\alpha = P/N$. When the loading rate is below some threshold, there exists at least a set of synaptic weights as a solution to correctly classify all the patterns. However, as α exceeds the threshold, it is impossible to find a compatible configuration of weights for all patterns [11]. This threshold is also defined as the storage capacity. Naturally, we define the energy of this model as the number of misclassified patterns as follows:

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} \Theta\left(-\frac{\sigma_0^\mu}{\sqrt{N}}\sum_i J_i\xi_i^\mu\right), \tag{1}$$

where $\Theta(x)$ is a step function with the convention that $\Theta(x) = 0$ if $x \leqslant 0$ and $\Theta(x) = 1$ otherwise. The prefactor $1/\sqrt{N}$ ensures that the statistical mechanics analysis leads to extensive free energy.

In the zero-temperature limit, the flat measure over the weights realizing the pattern-label associations can be computed as

$$P(\mathbf{J}) = \frac{1}{\mathrm{Z}}\prod_\mu \Theta\left(\frac{\sigma_0^\mu}{\sqrt{N}}\sum_i J_i\xi_i^\mu\right), \tag{2}$$

where Z is not only the partition function but also the number of solutions for the learning problem. Equation (2) can be derived from the finite-temperature Boltzmann measure $P(\mathbf{J}) \propto e^{-\beta E(\mathbf{J})}$. Notice that there is a gauge transformation $\xi_i^\mu \to \xi_i^\mu\sigma_0^\mu$ to each pixel of the input patterns that does not affect the Boltzmann measure. We thus assume $\sigma_0^\mu = +1$ for all patterns in the following analysis.

III. MEAN-FIELD MESSAGE-PASSING EQUATIONS FOR LEARNING

The belief propagation (BP) algorithm is an iterative mean-field equation to calculate the marginal probabilities of the synaptic state by passing beliefs between two types of nodes (function nodes and variable nodes) [3]. In other words, the beliefs or cavity probabilities can be assumed to be messages, and thus the BP algorithm is actually a mean-field message-passing equation. Taking pattern-classification constraints as function nodes and synaptic weights as variable nodes, we obtain the iterative equations for learning as follows [22,31]:

$$m_{i\to\nu} = \tanh\left(\sum_{\mu\neq\nu} u_{\mu\to i}\right), \tag{3a}$$

$$u_{\mu\to i} = \frac{1}{2}\left[\ln H\left(-\frac{\frac{1}{\sqrt{N}}\xi_i^\mu + w_{\mu\to i}}{\sqrt{\sigma_{\mu\to i}}}\right) - \ln H\left(-\frac{-\frac{1}{\sqrt{N}}\xi_i^\mu + w_{\mu\to i}}{\sqrt{\sigma_{\mu\to i}}}\right)\right], \tag{3b}$$

$$w_{\mu\to i} = \frac{1}{\sqrt{N}}\sum_{j\neq i} m_{j\to\mu}\xi_j^\mu, \tag{3c}$$

$$\sigma_{\mu\to i} = \frac{1}{N}\sum_{j\neq i}\left(1 - m_{j\to\mu}^2\right), \tag{3d}$$

where $H(x) = \int_x^\infty Dz$, Dz is a Gaussian measure, and $m_{i\to\nu}$ is a cavity magnetization parameter to parametrize the cavity probability $P(J_i|\{\boldsymbol{\xi}^{\mu\neq\nu}\}) = (1 + m_{i\to\nu}J_i)/2$. $w_{\mu\to i}$ and $\sigma_{\mu\to i}$ represent the mean and variance of the Gaussian distribution of $U_{\mu\to i} \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i} J_j\xi_j^\mu$, respectively. We have applied the central limit theorem to the sum $U_{\mu\to i}$ of weakly correlated terms. This mean-field approximation must be cross-checked by numerical experiments. In the following analysis, we use μ, ν to indicate function nodes or pattern constraints and i, j to indicate the variable nodes.

We remark that Eqs. (3a)–(3d) can be combined with an iterative reinforcement to develop an efficient solver. The reinforcement is a kind of soft decimation, which progressively

enhances or weakens current local fields (a summation of cavity biases $u_{\mu\to i}$) with an increasing probability with iterations. The algorithm terminates once a solution is found. This procedure yields the algorithmic threshold $\alpha_{\rm alg} \simeq 0.72$ [22]. During the stochastic reinforcement, the BP iteration does not require convergence, despite a convergence guarantee below the storage capacity. The algorithmic threshold is later found to be below a large-deviation threshold $\alpha_{\rm LD} \simeq 0.77$, after which the subdominant dense clusters fragment into separate regions [12,32]. However, our current analysis is restricted to the original BP iteration [Eqs. (3a)–(3d)], rather than the dynamics of reinforced BP and the geometric landscape. It remains challenging to use our framework (without a lengthy replica computation) to derive the landscape geometry, which relies heavily on the replica formula. The following analysis may shed light on this important research line.

IV. TIME EVOLUTION OF MESSAGE DISTRIBUTIONS

In this section, we study the iteration dynamics of the belief propagation. In the large-N limit, $u_{\mu\to i}$ can be approximated by the first-order Taylor expansion

$$u_{\mu\to i} = \frac{\xi_i^\mu}{\sqrt{N\sigma_{\mu\to i}}} \frac{G\left(-\frac{w_{\mu\to i}}{\sqrt{\sigma_{\mu\to i}}}\right)}{H\left(-\frac{w_{\mu\to i}}{\sqrt{\sigma_{\mu\to i}}}\right)}, \tag{4}$$

where $G(x) = \exp(-x^2/2)/\sqrt{2\pi}$ and $H(x) \equiv \int_x^\infty Dz$ with the Gaussian measure $Dz \equiv G(z)dz$. At the iteration step t, the macroscopic distributions of messages $m^t_{l\to\mu}$ and $u^t_{\mu\to l}$ are given by

$$\pi_1^t(x) = \frac{1}{NP}\sum_{l=1}^{N}\sum_{\mu=1}^{P}\delta\left(x - m^t_{l\to\mu}\right), \tag{5a}$$

$$\pi_2^t(\hat{x}) = \frac{1}{NP}\sum_{l=1}^{N}\sum_{\mu=1}^{P}\delta\left(\hat{x} - u^t_{\mu\to l}\right). \tag{5b}$$

According to the Kabashima's method [33], the time evolution of $\pi_1^t(x)$ and $\pi_2^t(\hat{x})$ can be written down in an iterative form as follows:

$$\pi_1^{t+1}(x) = \int\prod_{\mu=1}^{P-1} d\hat{x}_\mu \pi_2^t(\hat{x}_\mu)\delta\left(x - \tanh\left(\sum_{\mu=1}^{P-1}\hat{x}_\mu\right)\right), \tag{6a}$$

$$\pi_2^t(\hat{x}) = \int\prod_{l=1}^{N-1} dx_l \pi_1^t(x_l)\left\langle\delta\left(\hat{x} - \frac{\xi^\mu}{\sqrt{N\sigma_\mu}}\frac{G(X_\mu)}{H(X_\mu)}\right)\right\rangle_{\boldsymbol{\xi}}, \tag{6b}$$

$$X_\mu \equiv -\frac{\sum_{l=1}^{N-1}\xi_l^\mu x_l/\sqrt{N}}{\sqrt{1-\sum_{l=1}^{N-1}x_l^2/N}} = -\frac{w_\mu}{\sqrt{\sigma_\mu}}, \tag{6c}$$

where $\langle\cdots\rangle$ represents the disorder average over $\boldsymbol{\xi}$. Note that ξ^μ is independent of the pattern entries in the sum of X_μ.

We then introduce an auxiliary field $h^t_{l\to\mu} = \sum_{\nu\neq\mu} u^t_{\nu\to l} = \tanh^{-1}(m^t_{l\to\mu})$ and its macroscopic distribution $\rho(h)$. More precisely,

$$\rho^t(h) = \frac{1}{NP}\sum_{l=1}^{N}\sum_{\mu=1}^{P}\delta\left(h - h^t_{l\to\mu}\right). \tag{7}$$

When P becomes infinite (e.g., $P \propto N$), due to the central limit theorem, the distribution of the auxiliary field can be regarded as a Gaussian distribution:

$$\rho^t(h) = \int\prod_{\mu=1}^{P-1} d\hat{x}_\mu \pi_2^t(\hat{x}_\mu)\delta\left(h - \sum_{\mu=1}^{P-1}\hat{x}_\mu\right) \approx \frac{1}{\sqrt{2\pi F^t}}\exp\left[-\frac{(h-E^t)^2}{2F^t}\right], \tag{8}$$

where E^t and F^t are the mean and variance of the Gaussian distribution $\rho(h)$, respectively. In fact, $E^t = 0$ because of the setting that ξ^μ takes ±1 with equal probability. With the expression of $\rho^t(h)$, we get $\pi_1^{t+1}(x) = \int dh\rho^t(h)\delta[x - \tanh(h)]$ for Eq. (6a). Plugging this expression into Eq. (6b) and using Eq. (8), we obtain a compact expression for the update of F^t as

$$F^{t+1} = \frac{\alpha}{1-Q^t}\int Dz\left(\frac{G\left(-\sqrt{\frac{Q^t}{1-Q^t}}z\right)}{H\left(-\sqrt{\frac{Q^t}{1-Q^t}}z\right)}\right)^2, \tag{9a}$$

$$Q^t = \int Dz\tanh^2(\sqrt{F^t}z). \tag{9b}$$

We leave the technical details of this derivation to Appendix A. Note that this result is exactly identical to the saddle-point equation under the replica symmetric assumption, which we shall briefly introduce in Sec. VI.

V. MICROSCOPIC INSTABILITY OF THE BP ITERATION

In this section, we turn to the analysis of the microscopic stability of the BP equations at a fixed point. Provided that a field fluctuation $\delta h^t_{l\to\nu}$ is introduced around the fixed point $m^t_{l\to\nu} = m_{l\to\nu}$, the time evolution of $\delta h^t_{l\to\nu}$ is computed as

$$\delta h^t_{l\to\nu} = \sum_{\mu\neq\nu}\delta u_{\mu\to l} = \sum_{\mu\neq\nu}\frac{\xi_l^\mu}{\sqrt{N}}[L\delta w_{\mu\to l} + K\delta\sigma_{\mu\to l}], \tag{10}$$

where

$$\delta w_{\mu\to l} \equiv \frac{1}{\sqrt{N}}\sum_{i\neq l}\xi_i^\mu\left(1-m^2_{i\to\mu}\right)\delta h_{i\to\mu}, \tag{11a}$$

$$\delta\sigma_{\mu\to l} \equiv -\frac{2}{N}\sum_{i\neq l}m_{i\to\mu}\left(1-m^2_{i\to\mu}\right)\delta h_{i\to\mu}, \tag{11b}$$

$$K \equiv \left(\frac{w^2_{\mu\to l}}{\sigma_{\mu\to l}} + \frac{w_{\mu\to l}}{\sqrt{\sigma_{\mu\to l}}}\frac{G(X_\mu)}{H(X_\mu)} - 1\right)\frac{G(X_\mu)}{H(X_\mu)}\frac{1}{2\sigma^{\frac{3}{2}}_{\mu\to l}}, \tag{11c}$$

$$L \equiv -\frac{1}{\sigma_{\mu\to l}}\left(\frac{w_{\mu\to l}}{\sqrt{\sigma_{\mu\to l}}}\frac{G(X_\mu)}{H(X_\mu)} + \frac{G^2(X_\mu)}{H^2(X_\mu)}\right). \tag{11d}$$

Note that on the right-hand side of Eqs. (11a)–(11d), all messages or perturbations refer to their values at a previous step $(t-1)$. In the following analysis (including Appendix A), we omit this time index. We then define the macroscopic distribution of $\delta h^t_{l\to\nu}$ as $f^t(y)$ [33]. Due to the central limit theorem,

$f^t(y)$ can be assumed to be a Gaussian form, i.e.,

$$f^t(y) = \frac{1}{NP}\sum_{l=1}^{N}\sum_{\mu=1}^{P}\delta\left(y - \delta h^t_{l\to\mu}\right) \approx \frac{1}{\sqrt{2\pi b^t}}\exp\left[-\frac{(y-a^t)^2}{2b^t}\right], \tag{12}$$

where a^t and b^t are the mean and variance of the distribution, respectively. The time evolution of $f^t(y)$ is provided by a functional equation as follows:

$$f^{t+1}(y) = \int\prod_{\mu=1}^{P}\prod_{l=1}^{N} dy_{l\to\mu} f^t(y_{l\to\mu})\left\langle\delta\left(y - \sum_{\mu=1}^{P-1}\frac{\xi^\mu}{\sqrt{N}}[L\delta w_\mu + K\delta\sigma_\mu]\right)\right\rangle_{\{x_{l\to\mu}\},\boldsymbol{\xi}}. \tag{13}$$

Following a similar spirit as before, a^t is zero. We thus only need to focus on the update of b^t. We provide details of the derivation of this update in Appendix A. We finally obtain

$$b^{t+1} = \frac{\alpha}{(1-Q^t)^2}\int Dz\left(\frac{G(Z)}{H(Z)}\right)^2\left(Z - \frac{G(Z)}{H(Z)}\right)^2\int Dz\frac{1}{\cosh^4(\sqrt{F^t}z)}b^t \equiv \gamma b^t, \tag{14}$$

where $Z \equiv \sqrt{Q^t/(1-Q^t)}z$. When $\gamma < 1$, b^t converges to zero after iteration, indicating that the initially introduced fluctuation of the auxiliary field will eventually vanish. In contrast, when $\gamma > 1$, b^t would grow with iteration, which implies that the fluctuation will be amplified, leading to the instability of the fixed point. Therefore Eq. (14) provides the critical condition of the instability with respect to the growth of b^t, i.e.,

$$\frac{\alpha}{(1-Q)^2}\int Dz\left(\frac{G(Z)}{H(Z)}\right)^2\left(Z - \frac{G(Z)}{H(Z)}\right)^2\int Dz\frac{1}{\cosh^4(\sqrt{F}z)} = 1. \tag{15}$$

VI. EQUILIBRIUM PROPERTIES VIA THE REPLICA TRICK

In this section, we apply the replica trick to analyze the equilibrium properties of the binary perceptron. In the thermodynamic limit, the free energy has the self-averaging property, i.e., the distribution of the free energy for different realizations of learning is peaked at the typical value. Thus we can compute the disorder average given by $-\beta f = \langle \ln \mathrm{Z}\rangle$, where the average is carried out with respect to independent and identically distributed (i.i.d) random patterns. In fact, this disorder average is very hard to compute. However, by introducing n replicas of the original learning system and then setting $n \to 0$, we can obtain the free energy of the system in a mathematically concise way [3]:

$$-\beta f = \lim_{n\to 0, N\to\infty}\frac{\ln\langle \mathrm{Z}^n\rangle}{nN} = \lim_{n\to 0}\frac{\ln e^{NF_{\max}}}{nN} = \lim_{n\to 0}\frac{F_{\max}}{n}. \tag{16}$$

Here, we are interested in the zero-temperature limit (focusing on ground states). Therefore Eq. (16) is actually the entropy counting the number of solutions to the perceptron learning. By introducing replicas (copies of the system), we transfer a direct intractable treating of complex interactions in learning to handling the overlap matrix of states, which can be tackled by physics approximations, e.g., the RS ansatz in which the overlap does not depend on the specific replica index (permutation symmetry). An intuitive picture is that the RS ansatz is consistent with the deltalike distribution of messages on each link of the factor graph, and the broadening of the distribution (under the message perturbation) leads to the mathematical instability of the saddle point. We will come back to this point at the end of Sec. VII.

To compute $\ln\langle \mathrm{Z}^n\rangle$, we introduce n replicated synaptic weight vectors $\mathbf{J}^a$ $(a = 1, \ldots, n)$ as follows:

$$\begin{aligned}\langle \mathrm{Z}^n\rangle &= \left\langle\sum_{\{\mathbf{J}^a\}}\prod_{a,\mu}\Theta\left(\frac{1}{\sqrt{N}}\sum_i J_i^a\xi_i^\mu\right)\right\rangle \\ &= \int\prod_{a<b}\frac{dq^{ab}d\hat{q}^{ab}}{2\pi \mathrm{i}/N}\exp\left[-N\sum_{a<b}q^{ab}\hat{q}^{ab} + N\alpha G_0(\{q^{ab}\}) + NG_1(\{\hat{q}^{ab}\})\right],\end{aligned} \tag{17}$$

where we have introduced the state overlap $q^{ab} = \frac{1}{N}\sum_i J_i^a J_i^b$ and its associated conjugated counterpart $\hat{q}^{ab}$. The expressions of $G_0(\{q^{ab}\})$ (energy term) and $G_1(\{\hat{q}^{ab}\})$ (entropy term) are given as follows [11,31]:

$$G_0(\{q^{ab}\}) = \ln\int\prod_a\frac{d\lambda^a}{2\pi}\int_0^\infty dt^a e^{\mathrm{i}\sum_a\lambda^a t^a - \sum_{a<b}q^{ab}\lambda^a\lambda^b - \frac{1}{2}\sum_a(\lambda^a)^2}, \tag{18a}$$

$$G_1(\{\hat{q}^{ab}\}) = \ln\sum_{\{J^a\}}e^{\sum_{a<b}\hat{q}^{ab}J^aJ^b}. \tag{18b}$$

Plugging Eq. (17) into Eq. (16), we get the entropy

$$s = \lim_{n\to 0} \frac{1}{n} \max \left[-\sum_{a<b} q^{ab}\hat{q}^{ab} + \alpha G_0(\{q^{ab}\}) + G_1(\{\hat{q}^{ab}\}) \right]. \tag{19}$$

Under the RS ansatz $q^{ab} = q$, $\hat{q}^{ab} = \hat{q}$ for $a \neq b$, the extremization of Eq. (19) gives rise to the following saddle-point equations:

$$q = \int Dz \tanh^2(\sqrt{\hat{q}}z), \tag{20a}$$

$$\hat{q} = \frac{\alpha}{1-q} \int Dz \left(\frac{G\left(-\sqrt{\frac{q}{1-q}}z\right)}{H\left(-\sqrt{\frac{q}{1-q}}z\right)} \right)^2. \tag{20b}$$

The technical details are given in Appendix B. These saddle-point equations are again identical to Eqs. (9a) and (9b) derived from the BP equation.

VII. INSTABILITY OF THE REPLICA SYMMETRIC SOLUTION

The stability of the RS solution requires that the eigenvalues of the Hessian matrix (the second-derivative matrix) of $F_{\max}$ must be negative. The sign of the eigenvalues of this matrix evaluated at the RS solution tells us all the information about the stability [34]. We first introduce η^{ab} and ϵ^{ab} as the fluctuations around the RS solution as

$$q^{ab} = q + \eta^{ab}, \tag{21a}$$

$$\hat{q}^{ab} = \hat{q} + \epsilon^{ab}. \tag{21b}$$

By taking the Taylor expansion, we obtain $\frac{1}{2}\Delta$ as the second-order terms of $F_{\max}$, where

$$\Delta \equiv \alpha \sum_{\alpha\beta,\gamma\delta} \frac{\partial^2 G_0}{\partial q^{\alpha\beta}\partial q^{\gamma\delta}}\bigg|_{\eta^{\alpha\beta},\eta^{\gamma\delta}=0} \eta^{\alpha\beta}\eta^{\gamma\delta} - \sum_{\alpha\beta,\gamma\delta} \eta^{\alpha\beta}\epsilon^{\gamma\delta} + \sum_{\alpha\beta,\gamma\delta} \frac{\partial^2 G_1}{\partial \hat{q}^{\alpha\beta}\partial \hat{q}^{\gamma\delta}}\bigg|_{\epsilon^{\alpha\beta},\epsilon^{\gamma\delta}=0} \epsilon^{\alpha\beta}\epsilon^{\gamma\delta}, \tag{22}$$

where the prefactor α in the first term is the loading rate, the superscript of the order parameters indicates the replica index, $G_0 = G_0(\{q^{\alpha\beta}\})$, and $G_1 = G_1(\{\hat{q}^{\alpha\beta}\})$. In other words, the Hessian matrix looks like

$$\begin{bmatrix} \alpha\mathbf{H}_0 & -\mathbf{I} \\ -\mathbf{I} & \mathbf{H}_1 \end{bmatrix} \tag{23}$$

composed of four $\frac{n(n-1)}{2} \times \frac{n(n-1)}{2}$ blocks. $H_0^{(\alpha\beta)(\gamma\delta)} = \frac{\partial G_0}{\partial q^{\alpha\beta}\partial q^{\gamma\delta}}$, $H_1^{(\alpha\beta)(\gamma\delta)} = \frac{\partial G_1}{\partial \hat{q}^{\alpha\beta}\partial \hat{q}^{\gamma\delta}}$, and $\mathbf{I}$ is an identity matrix.

Following the analysis of Gardner and Derrida [23], we first consider the problem of diagonalizing the matrices of $\mathbf{H}_0$ and $\mathbf{H}_1$ separately. We then use the symmetry structure with respect to permutation of replica indices. The associated eigenvectors can be divided into three types (see details in Appendixes C and D). The first type are symmetric for all indices. The second type are symmetric for all but one specific index, and the third type are symmetric for all but two specific indices. In the limit of $n \to 0$, the second type of eigenvectors coincides with the first type of eigenvectors. The first type of eigenvectors defines the longitudinal fluctuations within the RS subspace [23,35]. This stability is already guaranteed by optimizing the action $F_{\max}$. In other words, the sufficient condition for $\lambda_{1,2} < 0$ is equivalent to the saddle-point equation [35].

Therefore only the third type of eigenvectors leads to the instability of the RS solution. This type of eigenvectors corresponds to the instability that is able to take the stationary point outside the RS subspace, capturing the transverse fluctuations. Suppose that the eigenvalues of these eigenvectors for $\mathbf{H}_0$ and $\mathbf{H}_1$ are γ_1 and γ_2, respectively. The related eigenvalues that cause the instability of the RS solution are given by the two eigenvalues of the following matrix:

$$\begin{bmatrix} \alpha\gamma_1 & -1 \\ -1 & \gamma_2 \end{bmatrix}. \tag{24}$$

The sign of the determinant determines the stability of the RS solution, i.e., the RS solution is stable only when $\alpha\gamma_1\gamma_2 < 1$. When $\alpha \to 0$, the determinant of this matrix is given by $\alpha\gamma_1\gamma_2 - 1 = -1$, which means that the product of the eigenvalues is negative. Therefore, in this limit, the RS solution is correct as expected. When α increases above a critical value, the sign of the determinant changes, which means that one of these eigenvalues changes its sign, thereby breaking the stability of the RS solution.

According to the calculation details in Appendix C, we have

$$\gamma_1 = \frac{1}{(1-q)^2} \int Dz \left(\frac{G(Z)}{H(Z)} \right)^2 \left(Z - \frac{G(Z)}{H(Z)} \right)^2, \tag{25a}$$

$$\gamma_2 = \int Dz \frac{1}{\cosh^4(\sqrt{\hat{q}}z)}, \tag{25b}$$

where $Z = \sqrt{q/(1-q)}z$. Therefore the critical condition for the transition to replica symmetry breaking is specified by

$$\alpha\gamma_1\gamma_2 = \frac{\alpha}{(1-q)^2} \int Dz \left(\frac{G(Z)}{H(Z)} \right)^2 \left(Z - \frac{G(Z)}{H(Z)} \right)^2 \times \int Dz \frac{1}{\cosh^4(\sqrt{\hat{q}}z)} = 1. \tag{26}$$

We thus conclude that Eq. (26) is identical to Eq. (15), which suggests that the equivalence between the BP instability and the transition to replica symmetry breaking can be established in perceptron learning systems. The replica symmetry breaking captures a hierarchical organization of replicas. In physics, this actually corresponds to the decomposition of the Gibbs measure into (exponentially or subexponentially) many pure states [36].

We finally carry out a numerical simulation to check whether the theoretical instability coincides with that obtained by running the BP in specific instances. As shown in Fig. 2, we observe the theoretical prediction, namely, that the de Almeida–Thouless (AT) [34] loading rate α_{AT} matches well the numerical estimation. The theoretical prediction is computed by solving Eqs. (9a) and (9b) and Eq. (15). During simulations, we estimate the convergence proportion as the

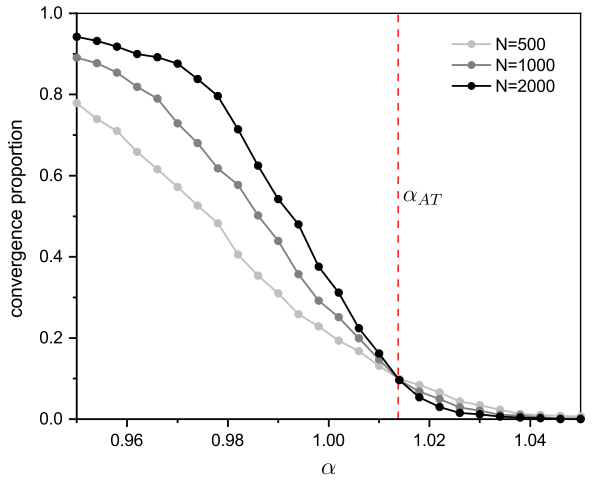

FIG. 2. The convergence proportion of the BP algorithm vs loading rate. The red dashed line marks the $\alpha_{\rm AT}$ computed by the stability condition equation of the RS solution. The three curves for different network sizes intersect at a point that coincides with $\alpha_{\rm AT}$. For each data point on the curves, we simulate M instances of binary perceptron. $M = 2000$ for $N = 500$; $M = 1000$ for $N = 1000$; and $M = 500$ for $N = 2000$.

fraction of instances for which the BP iteration converges within a prescribed criterion (e.g., all updated messages within a small deviation from the values at the previous iteration). It is expected from the plot that in the thermodynamic limit, the BP iteration does not converge beyond $\alpha_{\rm AT}$ with the probability tending to 1.

VIII. CONCLUSION

In this paper, from a physics perspective, we prove that the stability of the learning algorithm, derived using the physically intuitive cavity method, is connected to the stability of the replica symmetric saddle-point solution of the model. The equivalence between the physically intuitive cavity method and the mathematically concise replica method has also been explored in spin interaction systems [33], information transmission systems [37], linear estimation problems such as compressed sensing [38–40], and spectra estimation of random sparse matrices [41]. Our proof adds another piece of evidence of this equivalence in perceptron learning systems, by claiming rigorously (in the thermodynamic limit) the one-to-one correspondence between the BP instability and the AT instability of the equilibrium saddle point.

Our framework shows that the cumbersome replica analysis could be avoided in studying learning systems, e.g., the stability analysis considered in this paper. Therefore this work will hopefully inspire further studies of landscape analysis [12,28], unsupervised learning [7,42], and even deep learning, e.g., the current hot topic of learning in overparametrized neural networks [14].

ACKNOWLEDGMENT

We would like to thank other PMI members for discussions. This research was supported by National Natural Science Foundation of China Grants No. 12122515 and No. 11805284 (H.H.).

APPENDIX A: INSTABILITY ANALYSIS OF THE BP ITERATION

The iterative equation for the field distribution reads

$$\rho^{t+1}(h) = \int \prod_{\mu=1}^{P-1}\prod_{l=1}^{N-1} dh_{l\to\mu}\rho^t(h_{l\to\mu}) \times \left\langle \delta\left(h - \sum_{\mu=1}^{P-1}\frac{\xi^\mu}{\sqrt{N\sigma_\mu}}\frac{G(X_\mu)}{H(X_\mu)}\right)\right\rangle_{\xi}, \tag{A1a}$$

$$X_\mu = -\frac{\sum_{l=1}^{N-1}\xi_l^\mu \tanh(h_{l\to\mu})/\sqrt{N}}{\sqrt{1-\sum_{l=1}^{N-1}\tanh^2(h_{l\to\mu})/N}} = -\frac{w_\mu}{\sqrt{\sigma_\mu}}. \tag{A1b}$$

Notice that ξ^μ is independent of ξ_l^μ. We can thus calculate the average with respect to ξ^μ and ξ_l^μ separately. Due to the zero mean of ξ^μ, $E^t = \int h\rho^t(h)dh = 0$. Because $\xi_l^\mu = \pm 1$, we introduce a transformation $h_{l\to\mu} \to \xi_l^\mu h_{l\to\mu}$, which gives rise to

$$X_\mu = -\frac{\sum_{l=1}^{N-1}\tanh(h_{l\to\mu})/\sqrt{N}}{\sqrt{1-\sum_{l=1}^{N-1}\tanh^2(h_{l\to\mu})/N}}. \tag{A2}$$

Then the variance reads

$$\begin{aligned} F^{t+1} &= \int h^2\rho^{t+1}(h)dh = \int \prod_{\mu=1}^{P-1}\prod_{l=1}^{N-1} dh_{l\to\mu}\rho^t(h_{l\to\mu}) \\ &\quad\times \sum_{\mu=1}^{P-1}\left[\frac{1}{\sqrt{N\sigma_\mu}}\frac{G(X_\mu)}{H(X_\mu)}\right]^2 \\ &= \alpha\int\prod_{l=1}^{N-1} dh_l\rho^t(h_l)\left[\frac{1}{\sqrt{\sigma}}\frac{G(X)}{H(X)}\right]^2 \\ &= \alpha\mathbb{E}\left[\left(\frac{1}{\sqrt{\sigma}}\frac{G(X)}{H(X)}\right)^2\right], \end{aligned} \tag{A3}$$

where we have used the i.i.d. property of random patterns.

When $N \to \infty$, due to the law of large numbers, we have

$$\begin{aligned} \lim_{N\to\infty}\frac{1}{N}\sum_{l=1}^{N-1}\tanh^2(h_l) &= \mathbb{E}[\tanh^2(h_l)] \\ &= \int Dz\tanh^2(\sqrt{F^t}z) \equiv Q^t. \end{aligned} \tag{A4}$$

Due to the central limit theorem, we also have

$$\begin{aligned} \lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{l=1}^{N-1}\tanh(h_l) &= \sqrt{\mathbb{E}[\tanh^2(h_l)]}z + \mathbb{E}[\tanh(h_l)] \\ &= \sqrt{\mathbb{E}[\tanh^2(h_l)]}z = \sqrt{Q^t}z, \end{aligned} \tag{A5}$$

where $z \sim \mathcal{N}(0, 1)$. Therefore

$$\lim_{N\to\infty} X_\mu = -\frac{\lim_{N\to\infty}\sum_{l=1}^{N-1}\tanh(h_{l\to\mu})/\sqrt{N}}{\sqrt{1-\lim_{N\to\infty}\sum_{l=1}^{N-1}\tanh^2(h_{l\to\mu})/N}} = -\sqrt{\frac{Q^t}{1-Q^t}}z \equiv -Z. \tag{A6}$$

Plugging Eqs. (A4)–(A6) into Eq. (A3), we have

$$F^{t+1} = \frac{\alpha}{1-Q^t}\int Dz\left(\frac{G\left(-\sqrt{\frac{Q^t}{1-Q^t}}z\right)}{H\left(-\sqrt{\frac{Q^t}{1-Q^t}}z\right)}\right)^2. \tag{A7}$$

Next, we calculate the time evolution of a^t and b^t. Because of the zero mean of ξ^μ, it can also be proved that $a^t = 0$. In addition, b^{t+1} is the second-order moment of $f(y)$, i.e.,

$$\begin{aligned} b^{t+1} = \int y^2 f^{t+1}(y)dy &= \int \prod_{\mu=1}^{P}\prod_{l=1}^{N} dy_{l\to\mu} f^t(y_{l\to\mu})\left\langle\left(\sum_{\mu=1}^{P-1}\frac{\xi^\mu}{\sqrt{N}}[L\delta w_\mu + K\delta\sigma_\mu]\right)^2\right\rangle_{\{x_{l\to\mu}\},\boldsymbol{\xi}} \\ &= \frac{1}{N}\int \prod_{\mu=1}^{P}\prod_{l=1}^{N} dy_{l\to\mu} f^t(y_{l\to\mu})\sum_{\mu=1}^{P-1}\langle[L\delta w_\mu + K\delta\sigma_\mu]^2\rangle_{\{x_{l\to\mu}\},\boldsymbol{\xi}} \\ &= \frac{1}{N}\int \prod_{\mu=1}^{P}\prod_{l=1}^{N} dy_{l\to\mu} f^t(y_{l\to\mu})\sum_{\mu=1}^{P-1}\langle W_\mu\rangle_{\{x_{l\to\mu}\}}, \end{aligned} \tag{A8}$$

where

$$W_\mu \equiv \left\langle\left[\frac{L}{\sqrt{N}}\sum_{l=1}^{N-1}\xi_l^\mu\left(1-x_{l\to\mu}^2\right)y_{l\to\mu} - \frac{2K}{N}\sum_{l=1}^{N-1}x_{l\to\mu}\left(1-x_{l\to\mu}^2\right)y_{l\to\mu}\right]^2\right\rangle_{\boldsymbol{\xi}}. \tag{A9}$$

Performing the distribution-preserved transformation $x_{l\to\mu} \to \xi_l^\mu x_{l\to\mu}$ and neglecting the higher-order small terms in the large-N limit, we arrive at

$$\begin{aligned} W_\mu &= \left\langle\left[\frac{L}{\sqrt{N}}\sum_{l=1}^{N-1}\xi_l^\mu\left(1-x_{l\to\mu}^2\right)y_{l\to\mu} - \frac{2K}{N}\sum_{l=1}^{N-1}\xi_l^\mu x_{l\to\mu}\left(1-x_{l\to\mu}^2\right)y_{l\to\mu}\right]^2\right\rangle_{\boldsymbol{\xi}} \\ &\simeq \sum_{l=1}^{N-1}\left[\frac{L}{\sqrt{N}}\right]^2\left(1-x_{l\to\mu}^2\right)^2 y_{l\to\mu}^2, \end{aligned} \tag{A10}$$

and immediately we get

$$b^{t+1} = \lim_{N\to\infty} b^t\alpha\left\langle\sum_{l=1}^{N-1}\left[\frac{L}{\sqrt{N}}\right]^2\left(1-x_l^2\right)^2\right\rangle_{\{x_l\}}. \tag{A11}$$

Note that

$$L = -\lim_{N\to\infty}\frac{1}{\sigma}\left(\frac{w}{\sqrt{\sigma}}\frac{G(X)}{H(X)} + \frac{G^2(X)}{H^2(X)}\right) = -\frac{1}{(1-Q^t)}\left(\frac{G(-Z)}{H(-Z)}Z + \frac{G^2(-Z)}{H^2(-Z)}\right) \tag{A12}$$

and

$$\begin{aligned} \lim_{N\to\infty}\frac{1}{N}\sum_{l=1}^{N-1}\left(1-x_l^2\right)^2 &= 1 - 2\mathbb{E}\left[x_l^2\right] + \mathbb{E}\left[x_l^4\right] = 1 - 2\int Dz\tanh^2(\sqrt{F^t}z) + \int Dz\tanh^4(\sqrt{F^t}z) \\ &= \int Dz\frac{1}{\cosh^4(\sqrt{F^t}z)}. \end{aligned} \tag{A13}$$

Finally, we get

$$b^{t+1} = \frac{\alpha}{(1-Q^t)^2}\int Dz\left(\frac{G(Z)}{H(Z)}\right)^2\left(Z - \frac{G(Z)}{H(Z)}\right)^2\int Dz\frac{1}{\cosh^4(\sqrt{F^t}z)}b^t, \tag{A14}$$

where a statistically invariant change $z \to -z$ has been made.

APPENDIX B: DERIVATION OF SADDLE-POINT EQUATIONS

In this Appendix, we show explicitly how the replica computation is carried out. Applying the RS ansatz $q^{ab} = q$, $\hat{q}^{ab} = \hat{q}$ for $a \neq b$ to the energy term, we obtain

$$\begin{aligned}
G_0(q) &= \ln \int \prod_a \frac{d\lambda^a}{2\pi} \int_0^\infty dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}q(\sum_a \lambda^a)^2 - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2} \\
&= \ln \int Dz \int \prod_a \frac{d\lambda^a}{2\pi} \int_0^\infty dt^a e^{i\sum_a \lambda^a t^a - i\sum_a \lambda^a \sqrt{q}z - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2} \\
&= \ln \int Dz \int \prod_a \frac{d\lambda^a}{2\pi} \int_{-\sqrt{q}z}^\infty dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2} \\
&= \ln \int Dz \left[\int \frac{d\lambda}{2\pi} \int_{-\sqrt{q}z}^\infty dt e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2} \right]^n = \ln \int Dz \left[\int \frac{d\lambda}{2\pi} \int_{-\sqrt{\frac{q}{1-q}}z}^\infty dt e^{i\lambda t - \frac{1}{2}\lambda^2} \right]^n \\
&= \ln \int Dz \left[\int_{-\sqrt{\frac{q}{1-q}}z}^\infty Dt \right]^n = \ln \int Dz \left[H\left(-\sqrt{\frac{q}{1-q}}z \right) \right]^n,
\end{aligned} \tag{B1}$$

where we have rescaled $t = t\sqrt{1-q}$ and $\lambda = \lambda/\sqrt{1-q}$. Then we compute the entropy term $G_1(\hat{q})$ as

$$\begin{aligned}
G_1(\hat{q}) &= \ln \sum_{\{J^a\}} e^{\sum_{a<b} \hat{q}^{ab} J^a J^b} = \ln \sum_{\{J^a\}} e^{\hat{q}\sum_{a<b} J^a J^b} = \ln \sum_{\{J^a\}} e^{\frac{\hat{q}}{2}(\sum_a J^a)^2 - \frac{\hat{q}n}{2}} \\
&= \ln \int Dz \sum_{\{J^a\}} e^{\sqrt{\hat{q}}z\sum_a J^a - \frac{\hat{q}n}{2}} = \ln \int Dz e^{-\frac{\hat{q}n}{2}} \sum_{\{J^a\}} \prod_a e^{\sqrt{\hat{q}}zJ^a} \\
&= -\frac{\hat{q}n}{2} + \ln \int Dz \prod_a \left[\sum_{J^a} e^{\sqrt{\hat{q}}zJ^a} \right] = -\frac{\hat{q}n}{2} + \ln \int Dz [2\cosh\sqrt{\hat{q}}z]^n.
\end{aligned} \tag{B2}$$

Therefore the entropy of the model turns out to be

$$\begin{aligned}
s &= \lim_{n\to 0} \frac{1}{n} \max \left[-\frac{n(n-1)}{2} q\hat{q} - \frac{n}{2}\hat{q} + \ln \int Dz \left[H\left(-\sqrt{\frac{q}{1-q}}z \right) \right]^n + \ln \int Dz [2\cosh\sqrt{\hat{q}}z]^n \right] \\
&= \frac{q\hat{q}}{2} - \frac{\hat{q}}{2} + \int Dz \ln \left[H\left(-\sqrt{\frac{q}{1-q}}z \right) \right] + \int Dz \ln[2\cosh\sqrt{\hat{q}}z].
\end{aligned} \tag{B3}$$

Finally, we arrive at the saddle-point equations as follows:

$$\frac{\partial s}{\partial \hat{q}} = 0 \Rightarrow q = \int Dz \tanh^2(\sqrt{\hat{q}}z), \tag{B4}$$

$$\frac{\partial s}{\partial q} = 0 \Rightarrow \hat{q} = \frac{\alpha}{1-q} \int Dt \left(\frac{G\left(-\sqrt{\frac{q}{1-q}}t \right)}{H\left(-\sqrt{\frac{q}{1-q}}t \right)} \right)^2. \tag{B5}$$

APPENDIX C: INSTABILITY ANALYSIS OF THE RS SOLUTION

Considering the perturbation on the order parameters, we write the energy term G_0 as

$$\begin{aligned}
G_0 &= \ln \int \prod_a \frac{d\lambda^a}{2\pi} \int_0^\infty dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}q(\sum_a \lambda^a)^2 - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2 - \sum_{a<b} \eta^{ab}\lambda^a\lambda^b} \\
&= \ln \int Dz \int \prod_a \frac{d\lambda^a}{2\pi} \int_0^\infty dt^a e^{i\sum_a \lambda^a t^a - i\sum_a \lambda^a \sqrt{q}z - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2 - \sum_{a<b} \eta^{ab}\lambda^a\lambda^b} \\
&= \ln \int Dz \int \prod_a \frac{d\lambda^a}{2\pi} \int_{-\sqrt{q}z}^\infty dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2 - \sum_{a<b} \eta^{ab}\lambda^a\lambda^b},
\end{aligned} \tag{C1}$$

where we have shifted the integral variable $t^a \to t^a - \sqrt{q}z$. In addition, we define G_0' as

$$G_0' \equiv \int Dz \ln \int \prod_a \frac{d\lambda^a}{2\pi} \int_{-\sqrt{q}z}^{\infty} dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2 - \sum_{a<b}\eta^{ab}\lambda^a\lambda^b}. \tag{C2}$$

When $n \to 0$ and $\eta^{ab} \to 0$, we have

$$\begin{aligned}\lim_{\eta^{ab}\to 0}\lim_{n\to 0}\frac{G_0}{G_0'} &= \lim_{n\to 0}\frac{\ln \int Dz \int \prod_a \frac{d\lambda^a}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2}}{\int Dz \ln \int \prod_a \frac{d\lambda^a}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a (\lambda^a)^2}} \\ &= \lim_{n\to 0}\frac{\ln \int Dz\big[\int \frac{d\lambda}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2}\big]^n}{n\int Dz \ln\big[\int \frac{d\lambda}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2}\big]} = 1. \end{aligned} \tag{C3}$$

Therefore we can replace G_0 by G_0' in the above two limits in Eq. (22). We then get

$$H_0^{(\alpha\beta)(\gamma\delta)} \equiv \left.\frac{\partial^2 G_0}{\partial q^{\alpha\beta}\partial q^{\gamma\delta}}\right|_{\eta^{\alpha\beta},\eta^{\gamma\delta}=0} = \left.\frac{\partial^2 G_0'}{\partial \eta^{\alpha\beta}\partial \eta^{\gamma\delta}}\right|_{\eta^{\alpha\beta},\eta^{\gamma\delta}=0} = \langle \lambda^\alpha\lambda^\beta\lambda^\gamma\lambda^\delta\rangle - \langle\lambda^\alpha\lambda^\beta\rangle\langle\lambda^\gamma\lambda^\delta\rangle, \tag{C4}$$

where

$$\langle f(\lambda)\rangle \equiv \int Dz \frac{\int \prod_a \frac{d\lambda^a}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^a f(\lambda) e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a(\lambda^a)^2}}{\int \prod_a \frac{d\lambda^a}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a(\lambda^a)^2}}. \tag{C5}$$

At the RS saddle point, Eq. (C4) takes three possible values:

$$P = H_0^{(\alpha\beta)(\alpha\beta)} = \langle(\lambda^\alpha\lambda^\beta)^2\rangle - (\langle\lambda^\alpha\lambda^\beta\rangle)^2, \tag{C6a}$$

$$Q = H_0^{(\alpha\beta)(\alpha\gamma)} = \langle(\lambda^\alpha)^2\lambda^\beta\lambda^\gamma\rangle - \langle\lambda^\alpha\lambda^\beta\rangle\langle\lambda^\alpha\lambda^\gamma\rangle, \tag{C6b}$$

$$R = H_0^{(\alpha\beta)(\gamma\delta)} = \langle\lambda^\alpha\lambda^\beta\lambda^\gamma\lambda^\delta\rangle - \langle\lambda^\alpha\lambda^\beta\rangle\langle\lambda^\gamma\lambda^\delta\rangle, \tag{C6c}$$

where α, β, γ, and δ are not equal to each other. We then compute the relevant moment terms as follows:

$$\begin{aligned}\langle(\lambda^\alpha\lambda^\beta)^2\rangle &= \int Dz \frac{\int \prod_a \frac{d\lambda^a}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^a (\lambda^\alpha)^2(\lambda^\beta)^2 e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a(\lambda^a)^2}}{\int \prod_a \frac{d\lambda^a}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^a e^{i\sum_a \lambda^a t^a - \frac{1}{2}(1-q)\sum_a(\lambda^a)^2}} \\ &= \int Dz \frac{\int \frac{d\lambda^\alpha d\lambda^\beta}{(2\pi)^2}\int_{-\sqrt{q}z}^{\infty} dt^\alpha dt^\beta (\lambda^\alpha)^2(\lambda^\beta)^2 e^{i(\lambda^\alpha t^\alpha + \lambda^\beta t^\beta) - \frac{1}{2}(1-q)((\lambda^\alpha)^2+(\lambda^\beta)^2)}}{\int \frac{d\lambda^\alpha d\lambda^\beta}{(2\pi)^2}\int_{-\sqrt{q}z}^{\infty} dt^\alpha dt^\beta e^{i(\lambda^\alpha t^\alpha + \lambda^\beta t^\beta) - \frac{1}{2}(1-q)((\lambda^\alpha)^2+(\lambda^\beta)^2)}} \\ &= \int Dz \frac{\int \frac{d\lambda^\alpha}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^\alpha (\lambda^\alpha)^2 e^{i(\lambda^\alpha t^\alpha) - \frac{1}{2}(1-q)(\lambda^\alpha)^2}}{\int \frac{d\lambda^\alpha}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^\alpha e^{i(\lambda^\alpha t^\alpha) - \frac{1}{2}(1-q)(\lambda^\alpha)^2}} \frac{\int \frac{d\lambda^\beta}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^\beta (\lambda^\beta)^2 e^{i(\lambda^\beta t^\beta) - \frac{1}{2}(1-q)(\lambda^\beta)^2}}{\int \frac{d\lambda^\beta}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt^\beta e^{i(\lambda^\beta t^\beta) - \frac{1}{2}(1-q)(\lambda^\beta)^2}} \\ &= \int Dz \left(\frac{\int \frac{d\lambda}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt \lambda^2 e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2}}{\int \frac{d\lambda}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2}}\right)^2 \\ &= \int Dz(\overline{\lambda^2}[z])^2, \end{aligned} \tag{C7}$$

where $\overline{f(\lambda)}[z]$ is defined as

$$\overline{f(\lambda)}[z] = \frac{\int \frac{d\lambda}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt f(\lambda) e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2}}{\int \frac{d\lambda}{2\pi}\int_{-\sqrt{q}z}^{\infty} dt e^{i\lambda t - \frac{1}{2}(1-q)\lambda^2}}. \tag{C8}$$

Analogously, we obtain

$$\langle(\lambda^\alpha)^2\lambda^\beta\lambda^\gamma\rangle = \int Dz(\overline{\lambda}[z])^2\overline{\lambda^2}[z], \tag{C9}$$

$$\langle\lambda^\alpha\lambda^\beta\lambda^\gamma\lambda^\delta\rangle = \int Dz(\overline{\lambda}[z])^4, \tag{C10}$$

where $\overline{\lambda^2}[z]$ and $\overline{\lambda}[z]$ are computed as

$$\overline{\lambda}[z] = \frac{i}{\sqrt{1-q}}\frac{G(-Z)}{H(-Z)}, \tag{C11}$$

$$\overline{\lambda^2}[z] = \frac{1}{1-q}\frac{G(-Z)}{H(-Z)}Z, \tag{C12}$$

where $Z = \sqrt{q/(1-q)}z$.

Finally, we obtain γ_1 as

$$\gamma_1 = P - 2Q + R = \int Dz(\overline{\lambda^2}[z] - (\overline{\lambda}[z])^2)^2$$
$$= \frac{1}{(1-q)^2}\int Dz\left(\frac{G(Z)}{H(Z)}\right)^2\left(Z - \frac{G(Z)}{H(Z)}\right)^2. \tag{C13}$$

To compute γ_2, we first define

$$P' = H_1^{(\alpha\beta)(\alpha\beta)} = 1 - (\langle J^\alpha J^\beta\rangle)^2, \tag{C14a}$$

$$Q' = H_1^{(\alpha\beta)(\alpha\gamma)} = \langle J^\beta J^\gamma\rangle - \langle J^\alpha J^\beta\rangle\langle J^\alpha J^\gamma\rangle, \tag{C14b}$$

$$R' = H_1^{(\alpha\beta)(\gamma\delta)} = \langle J^\alpha J^\beta J^\gamma J^\delta\rangle - \langle J^\alpha J^\beta\rangle\langle J^\gamma J^\delta\rangle, \tag{C14c}$$

where $\langle f(J)\rangle$ is defined as

$$\langle f(J)\rangle = \int Dz\frac{\sum_{\{J^a\}} f(J)e^{\sqrt{\hat{q}}z\sum_a J^a}}{\sum_{\{J^a\}} e^{\sqrt{\hat{q}}z\sum_a J^a}}. \tag{C15}$$

We finally get

$$\gamma_2 = P' - 2Q' + R' = 1 - 2\langle J^\beta J^\gamma\rangle + \langle J^\alpha J^\beta J^\gamma J^\delta\rangle$$
$$= 1 - 2\int Dz\tanh^2(\sqrt{\hat{q}}z) + \int Dz\tanh^4(\sqrt{\hat{q}}z)$$
$$= \int Dz\frac{1}{\cosh^4(\sqrt{\hat{q}}z)}. \tag{C16}$$

APPENDIX D: EIGENVALUES OF THE HESSIAN MATRIX

Due to the symmetry with respect to permutation of replica indices, there are three types of eigenvectors for the Hessian matrix $\mathbf{H} = \{H^{(\alpha\beta)(\gamma\delta)}\}$ [34]. The first type of eigenvectors $\boldsymbol{\mu}_1$ has the following form:

$$\mu^{\alpha\beta} = a \quad \forall \quad \alpha < \beta. \tag{D1}$$

For all rows of $\mathbf{H}\boldsymbol{\mu}_1 = \lambda_1\boldsymbol{\mu}_1$, the equations can be generally written as

$$Pa + 2(n-2)Qa + \tfrac{1}{2}(n-2)(n-3)Ra = \lambda_1 a. \tag{D2}$$

While $n \to 0$, we obtain

$$\lambda_1 = P - 4Q + 3R. \tag{D3}$$

The second type of eigenvectors $\boldsymbol{\mu}_2$ has the following form:

$$\mu^{\alpha\theta} = \mu^{\theta\beta} = b, \quad \alpha, \beta \neq \theta,$$
$$\mu^{\alpha\beta} = c, \quad \alpha, \beta \neq \theta, \tag{D4}$$

where θ is the specific replica index. From $\mathbf{H}\boldsymbol{\mu}_2 = \lambda_2\boldsymbol{\mu}_2$, we obtain

$$Pb + (n-2)Qb + (n-2)Qc + \tfrac{1}{2}(n-2)(n-3)Rc = \lambda_2 b. \tag{D5}$$

Because $\mathbf{H}$ is a symmetric matrix, the eigenvectors corresponding to different eigenvalues are orthogonal to each other. Therefore $\boldsymbol{\mu}_1$ should be orthogonal to $\boldsymbol{\mu}_2$, leading to the following equation:

$$(n-1)ab + \tfrac{1}{2}(n-2)(n-1)ac = 0. \tag{D6}$$

Using Eq. (D5) and Eq. (D6) and setting $n \to 0$, we get

$$\lambda_2 = P - 4Q + 3R. \tag{D7}$$

Due to the choice of one specific replica, this eigenvalue is $(n-1)$-fold degenerate. We thus conclude that the eigenvalues of these two types of eigenvectors are the same in the limit of $n \to 0$. In fact, when $n \to 0$, c will also converge to b, making the forms of the two types of eigenvectors the same.

The third type of eigenvectors $\boldsymbol{\mu}_3$ has the following form:

$$\mu^{\theta\nu} = d, \quad \mu^{\alpha\nu} = \mu^{\nu\beta} = \mu^{\alpha\theta} = \mu^{\theta\beta} = e, \quad \alpha, \beta \neq \theta, \nu,$$
$$\mu^{\alpha\beta} = f, \quad \alpha, \beta \neq \theta, \nu, \tag{D8}$$

where θ and ν are the two specific replica indices. From $\mathbf{H}\boldsymbol{\mu}_3 = \lambda_3\boldsymbol{\mu}_3$, we obtain

$$Pd + 2(n-2)Qe + \tfrac{1}{2}(n-2)(n-3)Rf = \lambda_3 d. \tag{D9}$$

The orthogonality property is given by

$$da + 2(n-2)ea + \tfrac{1}{2}(n-2)(n-3)fa = 0, \tag{D10}$$

$$db + (n-2)eb + (n-2)ec + \tfrac{1}{2}(n-2)(n-3)fc = 0. \tag{D11}$$

Using Eqs. (D9)–(D11) and setting $n \to 0$, we get the $\frac{n(n-3)}{2}$-fold degenerate eigenvalue

$$\lambda_3 = P - 2Q + R. \tag{D12}$$

The total degeneracy (the number of linearly independent eigenvectors) of these three types of eigenvectors is $n(n-1)/2$, which implies that we have exhausted all the eigenvalues.

[1] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[2] L. Zdeborova, Understanding deep learning is also a job for physicists, Nat. Phys. **16**, 602 (2020).

[3] H. Huang, *Statistical Mechanics of Neural Networks* (Springer, Singapore, 2022).

[4] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks, Phys. Rev. Lett. **55**, 1530 (1985).

[5] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, Phys. Rev. A **45**, 4146 (1992).

[6] A. Barra, G. Genovese, P. Sollich, and D. Tantari, Phase diagram of restricted Boltzmann machines and generalized

Hopfield networks with arbitrary priors, Phys. Rev. E **97**, 022310 (2018).
[7] T. Hou and H. Huang, Statistical Physics of Unsupervised Learning with Prior Knowledge in Neural Networks, Phys. Rev. Lett. **124**, 248302 (2020).
[8] E. Gardner, Maximum storage capacity in neural networks, Europhys. Lett. **4**, 481 (1987).
[9] E. Gardner, The space of interactions in neural network models, J. Phys. A: Math. Gen. **21**, 257 (1988).
[10] E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, J. Phys. A: Math. Gen. **22**, 1983 (1989).
[11] W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, J. Phys. (Paris) **50**, 3057 (1989).
[12] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses, Phys. Rev. Lett. **115**, 128101 (2015).
[13] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, Proc. Natl. Acad. Sci. USA **113**, E7655 (2016).
[14] C. Baldassi, C. Lauditi, Enrico M. Malatesta, R. Pacelli, G. Perugini, and R. Zecchina, Learning through atypical "phase transitions" in overparameterized neural networks, arXiv:2110.00683.
[15] G. Parisi, Infinite Number of Order Parameters for Spin-Glasses, Phys. Rev. Lett. **43**, 1754 (1979).
[16] G. Parisi, A sequence of approximated solutions to the S-K model for spin glasses, J. Phys. A: Math. Gen. **13**, L115 (1980).
[17] H. Huang, Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses, J. Stat. Mech. (2017) 053302.
[18] M. Mézard, The space of interactions in neural networks: Gardner's computation with the cavity method, J. Phys. A: Math. Gen. **22**, 2181 (1989).
[19] M. Mézard and G. Parisi, The Bethe lattice spin glass revisited, Eur. Phys. J. B **20**, 217 (2001).
[20] M. Mézard and G. Parisi, The cavity method at zero temperature, J. Stat. Phys **111**, 1 (2003).
[21] J. S. Yedidia, W. T. Freeman, and Y. Weiss, Constructing free energy approximations and generalized belief propagation algorithms, IEEE Trans. Inf. Theory **51**, 2282 (2005).
[22] A. Braunstein and R. Zecchina, Learning by Message Passing in Networks of Discrete Synapses, Phys. Rev. Lett. **96**, 030201 (2006).
[23] E. Gardner and B. Derrida, Optimal storage properties of neural network models, J. Phys. A: Math. Gen. **21**, 271 (1988).
[24] H. Horner, Dynamics of learning for the binary perceptron problem, Z. Phys. B: Condens. Matter **86**, 291 (1992).
[25] H. K. Patel, Computational complexity, learning rules and storage capacities: A Monte Carlo study for the binary perceptron, Z. Phys. B: Condens. Matter **91**, 257 (1993).
[26] H. Huang and H. Zhou, Learning by random walks in the weight space of the Ising perceptron, J. Stat. Mech. (2010) P08014.
[27] H. Huang and H. Zhou, Combined local search strategy for learning in networks of binary synapses, Europhys. Lett. **96**, 58003 (2011).
[28] H. Huang and Y. Kabashima, Origin of the computational hardness for learning with binary synapses, Phys. Rev. E **90**, 052813 (2014).
[29] E. Abbe, S. Li, and A. Sly, Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron, arXiv:2102.13069.
[30] C. Baldassi, F. Gerace, H. J. Kappen, C. Lucibello, L. Saglietti, E. Tartaglione, and R. Zecchina, Role of Synaptic Stochasticity in Training Low-Precision Neural Networks, Phys. Rev. Lett. **120**, 268103 (2018).
[31] H. Huang, K. Y. M. Wong, and Y. Kabashima, Entropy landscape of solutions in the binary perceptron problem, J. Phys. A: Math. Theor. **46**, 375002 (2013).
[32] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Local entropy as a measure for sampling solutions in constraint satisfaction problems, J. Stat. Mech. (2016) 23301.
[33] Y. Kabashima, Propagating beliefs in spin-glass models, J. Phys. Soc. Jpn. **72**, 1645 (2003).
[34] J. R. L. de Almeida and D. J. Thouless, Stability of the Sherrington-Kirkpatrick solution of a spin glass model, J. Phys. A: Math. Gen. **11**, 983 (1978).
[35] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).
[36] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
[37] Y. Kabashima, A CDMA multiuser detection algorithm on the basis of belief propagation, J. Phys. A: Math. Gen. **36**, 11111 (2003).
[38] D. Donoho, A. Maleki, and A. Montanari, Message passing algorithms for compressed sensing, Proc. Natl. Acad. Sci. USA **106**, 18914 (2009).
[39] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices, J. Stat. Mech. (2012) P08009.
[40] L. Zdeborova and F. Krzakala, Statistical physics of inference: thresholds and algorithms, Adv. Phys. **65**, 453 (2016).
[41] F. Slanina, Equivalence of replica and cavity methods for computing spectra of sparse random matrices, Phys. Rev. E **83**, 011118 (2011).
[42] T. Hou, K. Y. M. Wong, and H. Huang, Minimal model of permutation symmetry in unsupervised learning, J. Phys. A: Math. Theor. **52**, 414001 (2019).