Mean-field message-passing equations in the Hopfield model and its generalizations

Marc Mézard
Physics Department, Ecole Normale Supérieure, PSL Research University, Paris
(Received 26 August 2016; published 14 February 2017)

Motivated by recent progress in using restricted Boltzmann machines as preprocessing algorithms for deep neural network, we revisit the mean-field equations [belief-propagation and Thouless-Anderson Palmer (TAP) equations] in the best understood of such machines, namely the Hopfield model of neural networks, and we explicit how they can be used as iterative message-passing algorithms, providing a fast method to compute the local polarizations of neurons. In the "retrieval phase", where neurons polarize in the direction of one memorized pattern, we point out a major difference between the belief propagation and TAP equations: The set of belief propagation equations depends on the pattern which is retrieved, while one can use a unique set of TAP equations. This makes the latter method much better suited for applications in the learning process of restricted Boltzmann machines. In the case where the patterns memorized in the Hopfield model are not independent, but are correlated through a combinatorial structure, we show that the TAP equations have to be modified. This modification can be seen either as an alteration of the reaction term in TAP equations or, more interestingly, as the consequence of message passing on a graphical model with several hidden layers, where the number of hidden layers depends on the depth of the correlations in the memorized patterns. This layered structure is actually necessary when one deals with more general restricted Boltzmann machines.

DOI: 10.1103/PhysRevE.95.022117

I. INTRODUCTION

The interest in neural networks has been revived recently by a series of practical successes using "deep neural networks" to solve important and difficult problems in artificial intelligence, ranging from image segmentation to speech recognition (see Ref. [1] and references therein). The crucial learning phase in these applications is often started by using techniques for unsupervised learning, like restricted Boltzmann machines (RBM) [2] or autoencoders [3], in order to obtain a first set of synaptic weights that is then optimized in a supervised learning process using back-propagation.

The unsupervised learning in RBMs is an important problem. Its difficulty comes from the necessity to compute the correlation functions of a general spin systems. The correlation functions can be approximated by numerical methods like Monte Carlo, but this is rather time-consuming. Alternative methods use local estimates of the correlations [4,5] or those that can be deduced by message-passing algorithms based on iteration of local mean-field equations. This last approach, which was pioneered in Refs. [6,7], has received more attention recently [8–10], and it seems that these sophisticated message-passing algorithms can be quite useful in RBM learning. In recent years, message-passing has proved successful, both for analytical studies and for algorithm design, in several important problems of computer science, including error correcting codes (for a review see, for instance, Ref. [11]), constraint satisfaction problems (for a review, see, for instance, Ref. [12]), statistical inference (for a review, see, for instance, Ref. [13]), compressed sensing [14–18], or learning in perceptrons [19–21].

The aim of this paper is to revisit the mean-field equations and their use as a message-passing algorithm, in the Hopfield model of neural networks [22]. The Hopfield model, a model of binary neurons interacting by pairs, with synaptic weights chosen in such a way that the neurons tend to polarize spontaneously towards one of the memorized "patterns", can also be seen as a RBM. It is, in fact, one of the best-understood models of neural networks and of RBMs, and it provides an excellent starting point to understand the mean-field message-passing equations and their possible use as algorithms.

The present paper addresses four issues. The first one is the derivation of the various types of mean-field equations in the Hopfield model, the second one is their use as an algorithm, and the third one is an analysis of the mean-field equations in a generalized Hopfield model where patterns have a combinatoric type of correlation. The fourth one is the generalization of the whole approach to RBMs which are of a more general type than the Hopfield model.

It is useful to clarify the mean-field equations in the case of the Hopfield model because several forms of these equations exist, under various names and acronyms as follows: belief propagation (BP), relaxed belief propagation (rBP), Thouless-Anderson Palmer equations (TAP), approximate message passing (AMP), and generalized approximate message passing (GAMP).

We shall see that each version is useful: BP and rBP form the basis of the statistical analysis called the cavity method [23] (also known as state evolution or density evolution in the recent computer-science literature) which gives the phase diagram of this problem. They can be used to derive TAP equations [24], which are also called AMP equations in the recent computer science literature. TAP equations were originally derived in the Hopfield model in Ref. [23]. Through our derivation of TAP equations as simplifications of the general BP equations (related to the one done in Ref. [25]), we confirm the validity of these equations, in spite of previous claims by Refs. [26,27] that they were incorrect. All methods actually lead to the same TAP equations as Ref. [23].

An important point which is clarified in the present approach concerns the use of message-passing mean-field equations in the "retrieval phase" of the Hopfield model, the phase where the neurons polarize spontaneously in the direction of one of the stored patterns (and where the model can be used as an associative memory). In this phase, the

usual simplification of BP equations into rBP, which assumes that messages have a Gaussian distribution, is incorrect and one must treat separately some of the messages which are associated with the specific pattern where the polarization develops. The equivalent of the rBP equations, taking into account this modification (called rBP-M in the following), are structurally distinct from the usual rBP equations. However, this distinction disappears when one writes TAP equation. This makes the TAP equations much better suited for algorithmic applications.

The Hopfield model is a system of binary neurons (or Ising spins), with infinite range pairwise interactions. It is thus intimately related to the infinite range model of spin glasses of Sherrington and Kirkpatrick (SK) [28], but it differs from it in the detailed structure of the interactions between spins. Instead of being independent random variables, the coupling constants between the spins are built from a set of predetermined patterns that one wants to memorize. This structure leads to a modification of the Onsager reaction term in the TAP equations. Our derivation shows that this modification is easily understood by using a representation of the Hopfield model with two layers, a layer of visible neuron variables and a layer of hidden pattern variables. The exchange of messages between these two layers (in which the Hopfield model is seen as a RBM) precisely leads to the modification of the Onsager reaction term. We will show that this structure can actually be iterated. We define a modified Hopfield model where the patterns are not independent random variables, but they are built by combinations of more elementary objects, called features. In this case, we show that the TAP equations can be understood by a neural network with three layers, in which one adds, between the layer of visible neuron variables and the layer of hidden pattern variables, another layer of hidden feature variables. This spontaneous emergence of more hidden layers when one handles a more structured type of problem is interesting in itself: One might hope that it could lead to an explanation of the success of multilayered network and deep learning in practical tasks where the information certainly contains a deep hierarchy of combinatorial correlations.

We do not address here the full problem of learning in the Hopfield model or in RBMs. We only study the "direct" problem of determining the polarization of each neuron (from which one can deduce the pair correlations by using linear response). However, a good control of this direct problem is an essential ingredient of most unsupervised learning protocols.

The paper is organized as follows: Section II provides basic definitions of the Hopfield model and recalls its phase diagram. Section III derives the mean-field equations. It starts with the phases where there is no spontaneous polarization of the neurons and derives successively the BP equations, their rBP simplification using Gaussian messages, and, finally, the TAP (or AMP) equations. It then studies the modifications of these equations when one works in the retrieval phase. The consistency of the BP equations with the standard replica results (a consistency which had been disputed in Ref. [27]) is then explicitly shown.

Section IV explains how the mean-field equations can be turned into algorithms by iterating them with a careful update schedule. Section V studies a modified Hopfield model in which the patterns are no longer independent, but they are built as combinations of more elementary random variables. We work out the modification of BP and TAP equations in this case, using a representation of the problem with two layers of hidden variables on top of the layer of visible neurons. Section VI derives the message-passing algorithms obtained from mean-field equations (BP, rBP, and TAP) in a general model of RBM. Section VII provides some concluding remarks and perspectives for further studies.

II. THE HOPFIELD MODEL

A. Definitions

In the Hopfield model [22], neurons are modeled as N binary spins s_i, $i = 1, \ldots, N$, taking values in $\{\pm 1\}$. These spins interact by pairs, the energy of a spin configuration is

$$E = -\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j. \tag{1}$$

This is a spin-glass model where the coupling constants J_{ij} take a special form. Starting from P "patterns", which are spin configurations

$$\xi_i^\mu = \pm 1, \quad i \in \{1, \ldots N\}, \quad \mu \in \{1, \ldots P\}, \tag{2}$$

the coupling constants are defined as

$$J_{ij} = \frac{1}{N}\sum_{\mu} \xi_i^\mu \xi_j^\mu. \tag{3}$$

Given an instance defined by the set of couplings $J = \{J_{ij}\}$, the Boltzmann distribution of the spins, at inverse temperature β, is defined as

$$P_J(s) = \frac{1}{Z} e^{(\beta/2)\sum_{i,j} J_{ij} s_i s_j}. \tag{4}$$

Using a Gaussian transformation, the partition function Z can be rewritten as

$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu}{\sqrt{2\pi/\beta}} \times \exp\left[-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{\mu,i} \frac{\xi_i^\mu}{\sqrt{N}} s_i \lambda_\mu\right]. \tag{5}$$

This expression shows that the Hopfield model is also a model of N binary spins s_i and P continuous variables with a Gaussian measure, λ_μ, interacting through random couplings $\xi_i^\mu/\sqrt{N}$ which are independent identically distributed random variables taking values $\pm 1/\sqrt{N}$ with probability 1/2. This is nothing but a restricted Boltzmann machine in which the visible neurons are binary variables that interact with P hidden continuous variables with a Gaussian distribution. The variable λ_μ can be interpreted as the projection of the spin configuration on the pattern μ, as suggested by the identity relating its mean $\langle \lambda_\mu \rangle$ and the expectations values of the spins:

$$\langle \lambda_\mu \rangle = \frac{1}{\sqrt{N}} \sum_i \xi_i^\mu \langle s_i \rangle. \tag{6}$$

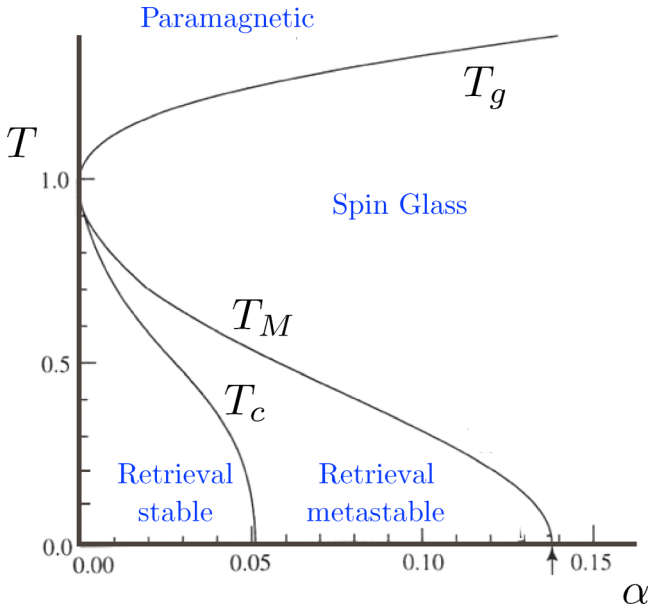

FIG. 1. Phase diagram of the Hopfield model from Ref. [30].

B. Known results

The phase diagram of the Hopfield model has been studied in detail in Refs. [29,30] and subsequent papers. In the thermodynamic limit where the number of neurons N and the number of patterns P go to infinity with a fixed ratio $\alpha = P/N$, the phase diagram is controlled by the temperature $T = 1/\beta$ and the ratio α. One finds three main phases:

(i) The paramagnetic phase. At high-enough temperatures, $T > T_g(\alpha)$, the spontaneous polarization of each neuron vanishes $\langle s_i \rangle = 0$.

(ii) The retrieval phase. In a regime of low-enough temperature and low-enough α, there exists a retrieval phase, where the neurons have a spontaneous polarization in the direction of one of the stored patterns μ. This means that, in the thermodynamic limit:

$$\frac{1}{N}\sum_i \langle s_i \rangle \xi_i^\mu = M, \tag{7}$$

$$\frac{1}{N}\sum_i \langle s_i \rangle \xi_i^\nu = 0 \ \ (\nu \neq \mu). \tag{8}$$

For symmetry reasons, there exist two retrieval states; One with a polarization $M > 0$ (where M is a function of α, β) and one with the polarization $-M$ (pointing opposite to the pattern).

The transition corresponding to the appearance of retrieval states is a first-order transition. One should thus distinguish two temperatures: At $T < T_M(\alpha)$, retrieval states first appear as metastable states, and at a lower temperature $T < T_c(\alpha)$, they become global minima of the free energy.

(iii) The spin-glass phase. In an intermediate range of temperature, or at large α, the neurons acquire a spontaneous polarization, but in some directions which are not in the direction of one of the patterns. In the spin-glass phase, the spin-glass order parameter q, defined by

$$q = \frac{1}{N}\sum_i \langle s_i \rangle^2 \tag{9}$$

is strictly positive, while

$$\forall \mu \ : \ \frac{1}{N}\sum_i \langle s_i \rangle \xi_i^\mu = 0. \tag{10}$$

The phase diagram is recalled in Fig. 1.

III. MEAN-FIELD EQUATIONS

A. Belief propagation

We use the representation (5) of the Hopfield model. Figure 2 shows the factor graph for this problem. The BP equations are written using the standard procedure, and we shall only sketch the derivation here and refer the reader to extensive presentations (see, for instance, Ref. [12]) for details. For the Hopfield model, this approach was first used by Ref. [25].

The main expressions used in the BP equations are as follows:

(i) The distribution of the neuron variable s_i in the absence of the pattern variable λ_μ. Denoted as $m_{i\to\mu}(s_i)$, this "cavity" distribution can be viewed as a message sent on the edge of the factor graph, from s_i towards λ_μ (see Fig. 2).

(ii) The distribution of the pattern variable λ_μ in the absence of the neuron variable s_i. This other cavity distribution is denoted as $m_{\mu\to i}(\lambda_\mu)$ and can be viewed as a message sent on the edge of the factor graph, from λ_μ towards s_i.

On top of the messages $m_{i\to\mu}(s_i)$ and $m_{\mu\to i}(\lambda_\mu)$, it is also convenient to introduce two auxiliary messages,

$$\hat{m}_{\mu\to i}(s_i) = \int \frac{d\lambda_\mu}{\sqrt{2\pi/\beta}}\, m_{\mu\to i}(\lambda_\mu) \times \exp\left[-(\beta/2)\lambda_\mu^2 + (\beta/\sqrt{N})\xi_i^\mu s_i \lambda_\mu\right], \tag{11}$$

$$\hat{m}_{i\to\mu}(\lambda_\mu) = \sum_{s_i} m_{i\to\mu}(s_i) \exp\left[(\beta/\sqrt{N})\xi_i^\mu s_i \lambda_\mu\right]. \tag{12}$$

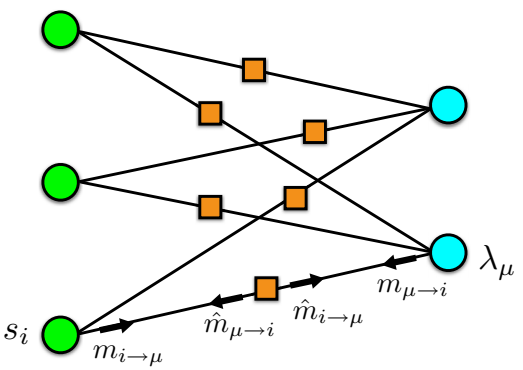

FIG. 2. Factor graph of the Hopfield model, in the representation using neuron variables (s_i, left circles) and pattern variables (λ_μ, right circles). For each pair of neuron variable and pattern variable, there is an interaction factor (squares). The graph also defines the messages used in belief propagation

The BP approximation amounts to assuming that, when one computes the message $m_{i\to\mu}(s_i)$, the contributions coming from each of the λ_ν variables are uncorrelated. This implies

$$m_{i\to\mu}(s_i) \cong \prod_{\nu(\neq\mu)} \hat{m}_{\nu\to i}(s_i) \tag{13}$$

[in this paper, the symbol "$\cong$" denotes equality up to a constant: If $p(\cdot)$ and $q(\cdot)$ are two measures on the same space—not necessarily normalized—we write $p(x) \cong q(x)$ if there exists $C > 0$ such that $p(x) = C\,q(x)$].

Similarly, when computing $m_{\mu\to i}(\lambda_\mu)$, one assumes that the contributions from each of the s_j variables are uncorrelated:

$$m_{\mu\to i}(\lambda_\mu) \cong e^{-(\beta/2)\lambda_\mu^2} \prod_{j(\neq i)} \hat{m}_{j\to\mu}(\lambda_\mu). \tag{14}$$

Equations (11)–(14) can be simplified as follows. The message $m_{i\to\mu}(s_i)$, being a probability of a binary variable, can be parameterized in terms of a single number $h_{i\to\mu}$, denoted "cavity field", defined by:

$$m_{i\to\mu}(s_i) \cong e^{\beta h_{i\to\mu} s_i}. \tag{15}$$

Similarly,

$$\hat{m}_{\mu\to i}(s_i) \cong e^{\beta \hat{h}_{\mu\to i} s_i}. \tag{16}$$

This gives the BP equations:

$$h_{i\to\mu} = \sum_{\nu(\neq\mu)} \hat{h}_{\nu\to i}, \tag{17}$$

$$e^{\beta \hat{h}_{\mu\to i} s_i} \cong \int d\lambda_\mu m_{\mu\to i}(\lambda_\mu) \exp\left[(\beta/\sqrt{N})\xi_i^\mu s_i \lambda_\mu\right], \tag{18}$$

$$\hat{m}_{i\to\mu}(\lambda_\mu) \cong \cosh\beta\left[h_{i\to\mu} + \left(\xi_i^\mu/\sqrt{N}\right)\lambda_\mu\right], \tag{19}$$

$$m_{\mu\to i}(\lambda_\mu) \cong e^{-(\beta/2)\lambda_\mu^2} \prod_{j(\neq i)} \hat{m}_{j\to\mu}(\lambda_\mu). \tag{20}$$

It is clear that in these equations one can use the explicit expressions of $\hat{m}_{i\to\mu}$ and $\hat{h}_{\mu\to i}$ and close the equations on the two sets of messages $h_{i\to\mu}$ and $m_{\mu\to i}$.

B. Relaxed BP equations

The general BP equations are not very useful because the messages $m_{\mu\to i}$ and $\hat{m}_{i\to\mu}$ are functions of the continuous variables λ_μ. However, the messages can be simplified by noticing that, in the thermodynamic limit, they are actually Gaussian distributions and can be parameterized by just two moments. This simplification leads to equations that are usually called "relaxed BP" (rBP) in the literature. It was first used in the cavity method for systems with long-range interactions [23] and has been developed in various problems of communication theory [31–34].

We shall first work out this simplification in the phase where there is no condensation on any pattern. Technically, this means that the distributions $m_{\mu\to i}(\lambda_\mu)$ are dominated by values of λ_μ which are finite (in the large N limit). It is easy to see that, in this case, (19) can be expanded around $\lambda_\mu/\sqrt{N} = 0$ and the BP equations (17)–(20) close under the hypothesis that the messages $m_{\mu\to i}(\lambda_\mu)$ are Gaussian distributions.

Under this assumption, one can parameterize these messages in terms of their two first moments. This leads to the so-called rBP equations. We define:

$$a_{\mu\to i} = \int d\lambda_\mu m_{\mu\to i}(\lambda_\mu)\lambda_\mu, \tag{21}$$

$$c_{\mu\to i} = \int d\lambda_\mu m_{\mu\to i}(\lambda_\mu)\lambda_\mu^2 - a_{\mu\to i}^2. \tag{22}$$

In order to derive the rBP equations, we first derive the asymptotic form of the function $\hat{m}_{i\to\mu}(\lambda_\mu)$ in the large-N limit:

$$\hat{m}_{i\to\mu}(\lambda_\mu) = \exp\left\{\beta\frac{\xi_i^\mu}{\sqrt{N}}\lambda_\mu \tanh(\beta h_{i\to\mu}) + \frac{\beta^2}{2N}\lambda_\mu^2[1-\tanh^2(\beta h_{i\to\mu})]\right\}. \tag{23}$$

Inserting this expression into (20) we get

$$c_{\mu\to i} = \frac{1}{\beta}\frac{1}{1-(\beta/N)\sum_{j(\neq i)}[1-\tanh^2(\beta h_{j\to\mu})]}, \tag{24}$$

$$a_{\mu\to i} = \frac{1}{\sqrt{N}}\frac{\sum_{j(\neq i)}\xi_j^\mu \tanh(\beta h_{j\to\mu})}{1-(\beta/N)\sum_{j(\neq i)}[1-\tanh^2(\beta h_{j\to\mu})]}. \tag{25}$$

Equation (17) and (18) can be rewritten as follows:

$$h_{i\to\mu} = \sum_{\nu(\neq\mu)} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i}. \tag{26}$$

Equations (25) and (26) form a set of $2NP$ equations for the $2NP$ variables $a_{\mu\to i}$ and $h_{i\to\mu}$. These are the rBP equations for the Hopfield model.

C. TAP (or AMP) equations

The rBP equations relate messages propagated along the edges of the factor graph (in the language of spin glasses they are "cavity equations"). In the large-N limit it is possible, and very useful for algorithmic purposes, to simplify these rBP equations and turn them into a set of equations which relate "site" quantities associated with the variable nodes of the factor graph. This allows us to go from $2NP$ variables to $N+P$ variables and leads to an effective reduction of computer time and memory. The equations that relate them are analogous to those found in Ref. [24] for spin glasses, hence the name TAP equations. In computer science, they are often called AMP equations [14,16–18,35–37]. To avoid confusion, notice that, in their paper on the Hopfield model [25], Kabashima and Saad use the same word (TAP equations) both for the rBP equations and what we call TAP equations. We prefer to use two different terms, in line with the terminology which is presently most common in information theory.

The site variables are local fields defined as follows:

$$H_i = \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i}, \tag{27}$$

$$A_\mu = \frac{1}{\sqrt{N}}\frac{\sum_j \xi_j^\mu \tanh(\beta h_{j\to\mu})}{1-(\beta/N)\sum_j[1-\tanh^2(\beta h_{j\to\mu})]}. \tag{28}$$

They give the expectation values of the variables

$$\langle s_i \rangle = M_i = \tanh(\beta H_i), \tag{29}$$

$$\langle \lambda_\mu \rangle = A_\mu. \tag{30}$$

We shall derive here a closed set of $N + P$ equations that relate these $N + P$ variables.

The main idea of the derivation of TAP equations comes from the observation that the rBP message $a_{\mu\to i}$ should be nearly equal to A_μ, up to small corrections that can be handled perturbatively in the large-N limit. Similarly, $h_{i\to\mu}$ is nearly equal to H_i, up to small corrections. Let us work out the explicit form of these corrections. We define

$$q = \frac{1}{N}\sum_i \tanh^2(\beta H_i). \tag{31}$$

We first notice that

$$h_{j\to\mu} = H_j - \frac{\xi_j^\mu}{\sqrt{N}} a_{\mu\to j}. \tag{32}$$

Therefore,

$$\frac{1}{N}\sum_j [1 - \tanh^2(\beta h_{j\to\mu})] \simeq 1 - q \tag{33}$$

up to corrections which vanish when $N \to \infty$. Therefore,

$$A_\mu \simeq \frac{1}{1-\beta(1-q)} \frac{1}{\sqrt{N}} \sum_j \xi_j^\mu \tanh(\beta h_{j\to\mu}) \tag{34}$$

and

$$a_{\mu\to i} \simeq A_\mu - \frac{1}{1-\beta(1-q)} \frac{1}{\sqrt{N}} \xi_i^\mu \tanh(\beta h_{i\to\mu}). \tag{35}$$

In this last expression, the second term is a correction of order $1/\sqrt{N}$. In this correction, we can substitute $h_{i\to\mu}$ by H_i, the difference would give a contribution of order $O(1/N)$ to $a_{\mu\to i}$, which can be neglected. Therefore,

$$a_{\mu\to i} \simeq A_\mu - \frac{1}{1-\beta(1-q)} \frac{1}{\sqrt{N}} \xi_i^\mu \tanh(\beta H_i). \tag{36}$$

Substituting this expression in the definition (27) of H_i we get

$$H_i \simeq \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} A_\nu - \frac{\alpha}{1-\beta(1-q)} \tanh(\beta H_i). \tag{37}$$

Considering now the definition (28) of A_μ, we can expand it as

$$\begin{aligned} A_\mu &= \frac{1}{1-\beta(1-q)} \sum_j \frac{\xi_j^\mu}{\sqrt{N}} \tanh\left[\beta\left(H_j - \frac{\xi_j^\mu}{\sqrt{N}} a_{\mu\to j}\right)\right] \\ &\simeq \frac{1}{1-\beta(1-q)} \sum_j \frac{\xi_j^\mu}{\sqrt{N}} \tanh\left[\beta\left(H_j - \frac{\xi_j^\mu}{\sqrt{N}} A_\mu\right)\right] \\ &\quad + O(1/N) \\ &\simeq \frac{1}{1-\beta(1-q)} \sum_j \frac{\xi_j^\mu}{\sqrt{N}} \\ &\quad \times \left[\tanh(\beta H_j) - \beta \frac{\xi_j^\mu}{\sqrt{N}} (1 - \tanh^2(\beta H_j) A_\mu\right]. \end{aligned} \tag{38}$$

This gives

$$A_\mu = \frac{1}{\sqrt{N}} \sum_j \xi_j^\mu \tanh(\beta H_j). \tag{39}$$

Equations (37) and (39), together with the definition (31), are the TAP (or AMP) equations which relate the $N + P$ variables H_i and A_μ. It turns out that they are linear in A_μ, and these variables can thus be eliminated (notice, however, that this is a specific feature of the Hopfield model, due to the Gaussian nature of variables λ_μ: As we shall see in Sec. VI, this is no longer true for more general RBMs, where the measure on λ_μ is non-Gaussian). Eliminating A_μ, we write closed equations for the N local fields H_i,

$$H_i = \frac{1}{N}\sum_j J_{ij} \tanh(\beta H_j) - \frac{\alpha}{1-\beta(1-q)} \tanh(\beta H_i). \tag{40}$$

An alternative presentation of these TAP equations are in terms of the local magnetizations $M_i = \tanh(\beta H_i)$:

$$M_i = \tanh\left[\beta \sum_j J_{ij} M_j - \frac{\alpha\beta}{1-\beta(1-q)} M_i\right], \tag{41}$$

$$= \tanh\left[\beta \sum_{j(\neq i)} J_{ij} M_j - \frac{\alpha\beta^2(1-q)}{1-\beta(1-q)} M_i\right], \tag{42}$$

$$q = \frac{1}{N}\sum_i M_i^2. \tag{43}$$

These TAP equations were first derived in Ref. [23] using the cavity method. The re-derivation that we have presented here uses a different approach, namely the BP equations and their simplification at large N, and obtains the same result.

The claims in Refs. [26,27], according to which these equations are wrong, were probably based on their misunderstanding of the presence of diagonal terms in (41). Actually, the TAP equations that they derive agree with ours, and with the original finding in Ref. [23], as can be seen explicitly in the form (42).

While Ref. [23] claimed (without writing the proof) that the TAP equations (43) reproduce the known equilibrium properties of the Hopfield model found with replicas, it was stated in Refs. [26,27] that they do not give the well-known value of the spin-glass transition temperature T_g and that they disagree with the result of the replica method. These statements are not correct. We provide below the explicit proof that our rBP and TAP equations are in perfect agreement with the replica result and therefore with the known value of T_g, as stated in Ref. [23]. The following derivation also gives a useful pedagogical example of how the equilibrium results can be obtained from the mean-field equations: The critical temperature can be analyzed through a study of the TAP equations, while the replica result for the order parameter can be obtained from a statistical analysis of the rBP equations.

D. rBP and TAP (AMP) equations in the retrieval phase

Let us work out the modifications that take place in the retrieval phase, when the measure condenses on one pattern (a similar analysis can be carried out easily in the mixed phase

where the condensation takes place on a finite number of patterns, we shall keep here to the retrieval phase). In the retrieval phase corresponding to pattern $\mu = 1$, one expects that the distribution of λ_1 will be dominated by values close to $\lambda_1 = M\sqrt{N}$. When deriving BP equations, the message $\hat{m}_{i\to 1}(\lambda_1)$ takes the form:

$$\hat{m}_{i\to 1}(\lambda_1 = M\sqrt{N}) = \cosh\beta\big(h_{i\to 1} + M\,\xi_i^1\big). \tag{44}$$

Therefore,

$$m_{1\to i}(\lambda_1 = M\sqrt{N}) \cong e^{N\psi_{1\to i}(M)}, \tag{45}$$

where

$$\psi_{1\to i}(M) = -\frac{\beta}{2}M^2 + \frac{1}{N}\sum_{j(\neq i)} \log\cosh\big[\beta\big(h_{j\to 1} + M\,\xi_j^1\big)\big]. \tag{46}$$

In the large-N limit, the measure $m_{1\to i}(M)$ is dominated by the maximum of the function $\psi_{1\to i}(M)$. One should notice that in the large-N limit this function converges to

$$\psi(M) = -\frac{\beta}{2}M^2 + \frac{1}{N}\sum_{j} \log\cosh\big[\beta\big(h_{j\to 1} + M\,\xi_j^1\big)\big]. \tag{47}$$

The maximum of $\psi(M)$ can be either in $M = 0$ or in $M = \pm M^*$, where M^* is the largest solution of the equation

$$M = \frac{1}{N}\sum_{j} \xi_j^1 \tanh\big[\beta\big(h_{j\to 1} + M\xi_j^1\big)\big]. \tag{48}$$

The retrieval phase is the phase where the maximum is obtained at $M = \pm M^*$. In this case, the rBP equations (25) and (26) are modified, because the messages $m_{1\to i}$, instead of being Gaussian distributions with finite means and variances, become dominated by values of λ_1 close to $M\sqrt{N}$. The new set of equations obtained in this regime will be denoted rBP-M (for relaxed belief propagation–magnetized) equations,

$$a_{\mu\to i} = \frac{1}{\sqrt{N}}\frac{\sum_{j(\neq i)}\xi_j^\mu \tanh(\beta h_{j\to\mu})}{1 - (\beta/N)\sum_{j(\neq i)}[1 - \tanh^2(\beta h_{j\to\mu})]}, \quad \mu \geqslant 2, \tag{49}$$

$$h_{i\to\mu} = \sum_{\nu(\neq\mu,1)} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i} + M\,\xi_i^1, \quad \mu \geqslant 2, \tag{50}$$

$$h_{i\to 1} = \sum_{\nu(\neq 1)} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i}. \tag{51}$$

The rBP-M equations in the retrieval phase with condensation on pattern 1 are given by (48)–(51).

It should be noticed that they involve a completely different estimate for the message $a_{1\to i}$ when compared to the rBP equations without condensation. In particular, they cannot be obtained from (25) and (26) by just assuming that $a_{1\to i}$ becomes of order $\sqrt{N}$ [such a procedure is unable to reproduce Eq. (48)]. The reason is that the condensation is a first-order transition, and the rBP equations in the retrieval phase correspond to a solution $M > 0$ to Eq. (48) that differs from the usual one with $M = 0$ [in which case one needs to consider the $O(1/\sqrt{N})$ corrections as in (25)]. The main drawback of these rBP-M equations is that one must use a different set of equations depending on the pattern towards which the system polarizes. This is quite inefficient for algorithmic applications: If one does not know *a priori* which pattern is being retrieved, one should run in parallel $P = \alpha N$ different algorithms, each one testing the possible polarization towards one of the patterns, and compare the results.

Fortunately, the situation is much better when considering TAP equations. It is straightforward to go from these rBP-M equations to the TAP (or AMP) equations for the retrieval phase. One gets:

$$H_i = \sum_{\nu\geqslant 2}\frac{\xi_i^\nu}{\sqrt{N}}A_\nu - \frac{\alpha}{1-\beta(1-q)}\tanh(\beta H_i) + M\,\xi_i^1, \tag{52}$$

$$A_\mu = \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu \tanh(\beta H_j), \quad \mu \geqslant 2, \tag{53}$$

$$M = \frac{1}{N}\sum_j \xi_j^1 \tanh(\beta H_j). \tag{54}$$

It turns out that these TAP equations are exactly the ones that would be obtained from the usual TAP equations (37) and (39), assuming that $A_1 = \sqrt{N}M$. This is rather remarkable considering the fact that the rBP-M equations in the retrieval phase cannot be obtained continuously from the rBP equations without retrieval (because of the first-order phase transition discussed above). The discontinuity in the set of rBP equations when going from the uncondensed to the retrieval phase thus disappears when one uses instead the TAP (GAMP) equations. This makes the TAP equations a much better choice for algorithmic applications.

E. Consistency with the replica results

While the critical temperature can be derived from TAP equations, a complete solution of the problem, including the computation of the spin-glass order parameter and the polarization, requires us to use the cavity method, which amounts here to a statistical analysis of the rBP (or rBP-M) equations.

1. Critical temperature

The paramagnetic solution of the TAP equations (41) and (43) is the solution with zero local magnetizations, $\forall i \;:\; M_i = 0$. The spin-glass transition is a second-order phase transition, and therefore its temperature $T_g = 1/\beta_g$ is the largest temperature where a solution with nonzero local magnetization exists. It can be found by linearizing the TAP equations (41) and (43) and identifying their instability point. Explicitly, the linearization gives:

$$M_i = \beta\sum_j J_{ij}M_j - \frac{\alpha\beta}{1-\beta}M_i + O(M^3). \tag{55}$$

The direction of instability is the one of the eigenvector of the J matrix with largest eigenvalue. Denoting by $\lambda_{\max}$ this largest eigenvalue, the value of β_g is given by

$$1 = \beta_g\lambda_{\max} - \frac{\alpha\beta_g}{1-\beta_g}. \tag{56}$$

By definition, $\lambda_{\max}$ is the largest eigenvalue of the matrix $N \times N$ matrix $J = (1/N)\xi\xi^T$, where the $N \times P$ matrix ξ has independent identically distributed random entries taking values ± 1 with probability $1/2$. In fact, the distribution of the largest eigenvalue of J concentrates around

$$\lambda_{\max} = (1 + \sqrt{\alpha})^2. \tag{57}$$

This result can be derived using the replica method or the cavity method. An easy way to obtain it is to realize that the value of $\lambda_{\max}$ depends only on the first two moments of the distribution of the matrix elements ξ_i^μ. In particular, it is the same as the one which would be obtained if the entries of ξ were independent identically distributed with a normal distribution of mean 0 and variance 1. This last case is very well known since the work of Marcenko and Pastur [38], and it gives the value of $\lambda_{\max}$ written in (57).

Using (57), the value of β_g obtained from (56) is

$$\beta_g = \frac{1}{1 + \sqrt{\alpha}}. \tag{58}$$

This agrees with the well-known result of Ref. [30] for the critical temperature: $T_g = 1 + \sqrt{\alpha}$.

2. *Order parameter*

The cavity or BP equations can be used in two distinct ways: On a single instance they can be solved by iteration, and if a fixed point is found, then this idea may be used as an algorithm for estimating the local magnetizations. But in the case where the instances are generated from an ensemble (like the case that we study here, where the ξ_i^μ are independent identically distributed random variables), one can also perform a statistical analysis of the equation. This is the essence of the cavity method and is also known in the literature on message passing as the density evolution.

We will show that this statistical analysis of the cavity equations gives the same results as the replica method, as claimed in Ref. [23], and contrary to the statements of Ref. [27]. For simplicity, we keep here to the "replica symmetric" approximation.

We start from the rBP equations. Considering first the equation (26) giving the cavity field $h_{i\to\mu}$, we notice that, as the variables ξ_i^μ are independent identically distributed, provided that the correlations of the messages $a_{\nu\to i}$ are small enough (this is the essence of the replica symmetric approximation; see Refs. [12,23]), the cavity field $h_{i\to\mu}$ has a Gaussian distribution with mean 0 and a variance which is independent of the indices i and μ and that we denote by $\overline{h^2}$. Similarly, $a_{\nu\to i}$ has a Gaussian distribution with mean 0 and a variance which is independent of the indices i and ν and that we denote by $\overline{a^2}$. The rBP equations (25) and (26) relate these two variances:

$$\overline{h^2} = \alpha \overline{a^2}, \tag{59}$$

$$\overline{a^2} = \frac{q}{[1 - \beta(1-q)]^2}. \tag{60}$$

We thus obtain:

$$q = \overline{\tanh^2(\beta h)} = \int \frac{dh}{\sqrt{2\pi\Phi}} e^{-h^2/(2\Phi)} \tanh^2(\beta h), \tag{61}$$

where

$$\Phi = \frac{\alpha q}{[1 - \beta(1-q)]^2}, \tag{62}$$

Equations (61) and (62) are exactly the well-known equations [30] that allow us to compute the spin-glass order parameter q in the spin-glass phase of the Hopfield model in the replica-symmetric framework.

In the retrieval phase, the same reasoning can be applied starting form the rBP-M equations (48) and (51). One finds:

$$q = \overline{\tanh^2(\beta h + \beta \xi M)}, \tag{63}$$

$$M = \overline{\xi \tanh(\beta h + \beta \xi M)}, \tag{64}$$

where the overline denotes the average with respect to the field h, which has a Gaussian distribution of variance Φ, and the binary variable ξ which takes values ± 1 with probability $1/2$. These are precisely the equations obtained in the retrieval phase with the replica method [30]. In particular, one can identify the appearance of the retrieval phase (the line T_M in Fig. 1) by analyzing when the equations (63) and (64) have a solution with $M \neq 0$ (in order to derive the value of the equilibrium phase transition T_c one needs to compute the free energy in the retrieval phase and in the spin-glass phase and see when they are equal).

IV. ALGORITHMS: ITERATIONS AND TIME INDICES

Mean-field equations are usually solved by iteration and interpreted as message-passing algorithms. Turning a set of mean-field equations into an iterative algorithm involves a certain degree of arbitrariness concerning the way the equations are written and the "time indices" concerning the update. A proper choice of time indices may result in an algorithm with much better convergence properties, as underlined, for instance, in Refs. [13,39]. Here we review the most natural choice for AMP iterations and their consequences.

A. rBP equations

The rBP equations (25) and (26) are usually iterated as follows:

$$a_{\mu\to i}^{t+1} = \frac{1}{\sqrt{N}} \frac{\sum_{j(\neq i)} \xi_j^\mu \tanh\left(\beta h_{j\to\mu}^t\right)}{1 - (\beta/N)\sum_{j(\neq i)}\left[1 - \tanh^2\left(\beta h_{j\to\mu}^t\right)\right]}, \tag{65}$$

$$h_{i\to\mu}^{t+2} = \sum_{\nu(\neq\mu)} \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i}^{t+1}. \tag{66}$$

There exist various types of update schemes. One can distinguish two main classes:

(a) In the parallel update, starting from a configuration of the h messages at time t, one computes all the a messages using (65). Then one computes all the new h messages at time $t + 2$ using (66), with the a messages of time $t + 1$ (therefore the h messages are defined at even times, and the a messages are defined at odd times). In two time steps, all the messages are updated.

(b) In an update in series, one picks up a message at random (or, better, one can use a random permutation of all messages to decide on the sequence of updates), and one updates it using

either (65)—if the message is an a message—or (66). In the case of random permutations, all messages are updated after $2NP$ time steps.

In the parallel update scheme, one can easily follow the evolution in time of the overlap q^t. Using (65) and (66), one can perform again the analysis of Sec. III E 2 keeping the time indices. This gives

$$q^{t+2} = \int \frac{dh}{\sqrt{2\pi\,\Phi^t}} e^{-h^2/(2\Phi^t)} \tanh^2(\beta h), \tag{67}$$

where

$$\Phi^t = \frac{\alpha q^t}{[1-\beta(1-q^t)]^2}. \tag{68}$$

It is easy to see that these equations converge (to $q = 0$) when $T > T_g = 1 + \sqrt{\alpha}$.

B. TAP equations

We can now repeat the previous derivation of the TAP equations, keeping track of the time indices that were written in the previous subsection. We keep here to the case of parallel update. Defining:

$$A_\mu^{t+1} = \frac{1}{\sqrt{N}} \frac{\sum_j \xi_j^\mu \tanh\left(\beta h_{j\to\mu}^t\right)}{1-(\beta/N)\sum_j\left[1-\tanh^2\left(\beta h_{j\to\mu}^t\right)\right]}, \tag{69}$$

$$H_i^{t+2} = \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} a_{\nu\to i}^{t+1}, \tag{70}$$

one gets

$$A_\mu^{t+1} = \frac{1}{1-\beta(1-q^t)} \frac{1}{\sqrt{N}} \sum_j \xi_j^\mu \tanh\left(\beta H_j^t\right) - \frac{\beta(1-q^t)}{1-\beta(1-q^t)} A_\mu^{t-1}, \tag{71}$$

$$H_i^{t+2} = \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} A_\nu^{t+1} - \frac{\alpha}{1-\beta(1-q^t)} \tanh\left(\beta H_i^t\right). \tag{72}$$

Equations (71) and (72) give the algorithmic version of TAP equations, used through a parallel iteration.

Again, the A variables can be eliminated from these equations, leaving the TAP equations written in terms of the local fields H_i^t or the magnetizations $M_i^t = \tanh(\beta H_i^t)$. This requires a little bit of care since the A variables appear with different time indices in the two sides of Eq. (71). Defining $u^t = \beta(1-q^t)$, we can re-express (71) as

$$A_\mu^{t+1} + \frac{u^t}{1-u^t} A_\mu^{t-1} = \frac{1}{1-u^t} \frac{1}{\sqrt{N}} \sum_j \xi_j^\mu \tanh\left(\beta H_j^t\right). \tag{73}$$

This suggest to use (72) at times $t+2$ and at time t in the form:

$$H_i^{t+2} + \frac{u^t}{1-u^t} H_i^t = \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}} \left(A_\nu^{t+1} + \frac{u^t}{1-u^t} A_\nu^{t-1} \right) - \frac{\alpha}{1-u^t} M_i^t - \frac{u^t}{1-u^t} \frac{\alpha}{1-u^{t-2}} M_i^{t-2}. \tag{74}$$

Substituting (74) into (73), one can eliminate the A variables. Using $\tau = t/2$ we get

$$H_i^{\tau+1} = \frac{1}{1-u^\tau} \left[\sum_j J_{ij} M_j^\tau - \alpha M_i^\tau - u^\tau H_i^\tau - \frac{\alpha u^\tau}{1-u^{\tau-1}} M_i^{\tau-1} \right]. \tag{75}$$

This final form of the iterative algorithm corresponding to TAP equation involves a kind of memory term (the polarization of neuron i at time $\tau+1$ is obtained from polarizations at time τ and $\tau-1$), a phenomenon that was first found in the context of TAP equations for the SK model [39] and used in Ref. [13].

This algorithm has many advantages. It involves only N fields, and therefore its iteration is fast, and above all it can develop a spontaneous polarization towards one of the stored patterns (while in the rBP equations one would need to use a different equation for each of the patterns).

C. Numerical results

The iteration of TAP equations (71) and (72), or, equivalently, their expression in terms of the H fields only (75), is a fast algorithm for solving the Hopfield model (in the sense of obtaining the local polarizations of the neuron variables). We have tested it in the retrieval phase, starting from a configuration with overlap M_0 with one randomly chosen pattern μ_0. This means that we generate an initial spin configuration s_i^0 as

$$s_i^0 = \xi_i^{\mu_0} \quad \texttt{with probability} \quad (1+M_0)/2, \tag{76}$$

$$s_i^0 = -\xi_i^{\mu_0} \quad \texttt{with probability} \quad (1-M_0)/2. \tag{77}$$

The initial field H_i^0 at time zero is then fixed as $H_i^0 = (8/\beta)s_i^0$, so $(1/N)\sum_i \xi_i^{\mu_0} \tanh(\beta H_i^0) = M_0$ up to fluctuations of order $1/\sqrt{N}$.

Figure 3 shows the probability that the iteration of these equations converges to a fixed point with a value of overlap with μ_0, given by $(1/N)\sum_i \tanh(\beta H_i)\xi_i^{\mu_0}$, larger than 0.95 [the convergence is defined by the fact that, in (75), the average value of $|H_i^{\tau+1} - H_i^\tau| < 10^{-6}$]. The simulations were carried out with networks of $N = 1000$ neurons. The maximal number of iterations was fixed to 200, but in practice we notice that when the algorithm converges it does so in a few iterations, of order 10 to 20.

It should be noticed that the iteration of simpler versions of the mean-field equations, either the naive mean-field equations or the SK-TAP equations with the correct time indices of Ref. [39], also converge when initialized in the same conditions. Actually, the basin of attraction for convergence to an overlap >0.95 with the pattern is larger for naive mean field than it is for SK-TAP, and the one for SK-TAP is larger than for the correct Hopfield TAP equations. This is probably due to the fact, noticed in Ref. [25], that the fixed point reached by naive mean field is actually closer to the pattern than the fixed point reached by SK-TAP, which is itself closer to the pattern than the one obtained by iterating TAP equations. However, the TAP equations have one major advantage: They give the

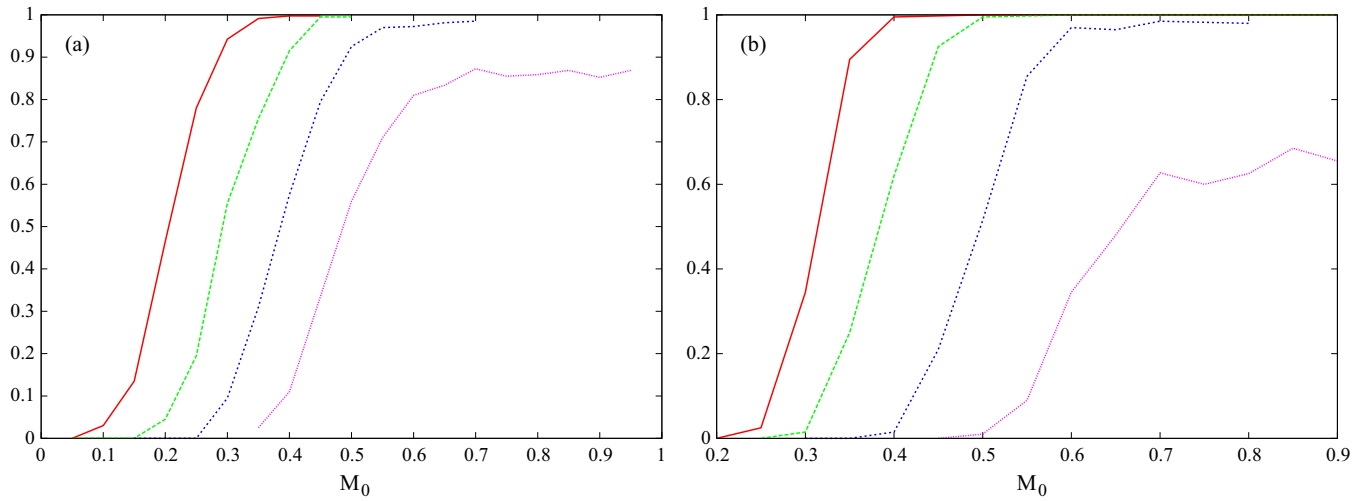

FIG. 3. Iteration of TAP equations using (75). The probability of convergence is plotted versus the overlap to a randomly chosen initial pattern μ_0. Left-hand figure: temperature $T = 0.01$, simulation with $N = 1000$ neurons, and $P = 40,60,80,100,120$ patterns (from left to right). Right-hand figure: temperature $T = 0.3$, simulation with $N = 1000$ neurons, and $P = 40,60,80,100$ patterns (from left to right).

values of the polarizations of neurons which become exact in the thermodynamic limit.

Although they are *a priori* much slower, one may also wonder about the numerical behavior of the rBP-M equations in the retrieval phase (48)–(51). These can be iterated for instance as follows. One first initializes the fields $h_{j\to\mu}$ to $(8/\beta)s_i^0$ as above. Then:

(i) Given the fields $\{h_{j\to\mu}\}$, compute the "magnetization" M by solving Eq. (48) for condensation on pattern μ_1. This equation can be solved for instance by iteration, starting from the value of M found at the previous iteration (initially we start the iteration with $M = 1$).

(ii) Compute the fields $\{a_{\mu\to j}\}$ using (49).

(iii) Compute the fields $\{h_{j\to\mu}\}$ using (50) and (51) (NB: in order to improve convergence, we use a "damping" term in the computation of the fields $\{h_{j\to\mu}\}$: Instead of substituting the old value of $h_{j\to\mu}$ by the new one, we substitute it by the arithmetic mean of the old and the new one).

Experimenting with these rBP-M equations we find that, in the retrieval phase, they converge fast (in a few iterations). However, the fixed point to which they converge depends very little on the initial overlap M_0 with pattern μ_0. Actually, the fixed point is the one corresponding to the pattern $\mu_1 = 1$ onto which the condensation has been assumed: The rBP-M equations tell us that each of the memorized patterns is a fixed point, and the initialization of the $h_{j\to\mu}$ messages in the direction of pattern μ_0 plays little role. Instead of using the pattern as an initial condition, the rBP-M equations should be used numerically with pattern μ_0 playing the role of a permanent external field that biases the activity of each neuron, as was done, for instance, in Ref. [40].

V. MODIFIED HOPFIELD MODEL: CORRELATED PATTERNS WITH COMBINATORIAL STRUCTURE

From its definition (1), the Hopfield model is a type of spin glass. It differs from the SK model by the structure of couplings. In the SK model, one draws each J_{ij} (for $i < j$) as an independent random variable with mean zero and variance $1/N$. In the Hopfield model, one builds the J_{ij} coupling constants as bilinear superposition of patterns, see (3). It turns out that this modification in the generation of the couplings induces a crucial modification of the TAP equations. In the SK model, the TAP equations are as follows [24]:

$$M_i = \tanh\left[\beta \sum_{j(\neq i)} J_{ij} M_j - \beta^2(1-q)M_i\right]. \tag{78}$$

The structure is the same in the Hopfield model, but the precise form of the second term (the so-called Onsager reaction term) differs. For an instructive comparison, it is useful to rescale the interactions of the Hopfield model in such a way that the variance of the couplings are $1/N$, defining thus $J_{ij} = \frac{1}{\sqrt{\alpha}}\frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu$. This simple rescaling can be absorbed in a rescaling of β, and our TAP equations (41) become, in this rescaled Hopfield model,

$$M_i = \tanh\left[\beta \sum_{j(\neq i)} J_{ij} M_j - \frac{\beta^2(1-q)}{1-\beta(1-q)/\sqrt{\alpha}} M_i\right]. \tag{79}$$

Therefore, the change of structure of the J_{ij} random variables leads to a modification of the TAP equations, where the Onsager term acquires a denominator $1/[1-\beta(1-q)/\sqrt{\alpha}]$. Clearly, in the large-$\alpha$ limit one recovers the TAP equations of the SK model, as it should be, since the correlations between the J_{ij} become irrelevant in this limit.

The fact that the TAP equations depend on the type of structure of the couplings J_{ij} poses a challenge for their use in practical applications, where one does not really know the structure of these couplings. One elegant way out consists in adapting the reaction term to the concrete set of couplings to which one is applying the method [7,41]. Our approach in the present paper considers instead an alternative representation of the Hopfield model, in which the visible neuron variables interact with a hidden layer of pattern variables. In this expanded representation, the couplings between the visible

and the hidden units are nothing but the patterns, which are independent random variables. Therefore the message passing equations (BP, rBP, and eventually TAP) can be written safely and give the result.

The standard results of the Hopfield model hold as long as the ξ_i^μ are independent identically distributed random variables with zero mean and unit variance. We test our approach by studying a generalization of the Hopfield model in which the patterns are no longer independent random variables. We shall study the case where the patterns have a correlation, created from the following structure:

$$\xi_i^\mu = \frac{1}{\sqrt{\gamma N}} \sum_{r=1}^{\gamma N} u_i^r v_\mu^r, \tag{80}$$

where the u_i^r are independent identically distributed random variables drawn from a distribution P_u with zero mean and unit variance, and the v_μ^r are independent identically distributed random variables drawn from a distribution P_v with zero mean and unit variance. Note that the scaling has been chosen such that, in the large-γ limit, the pattern elements ξ_i^μ become independent identically distributed Gaussian random variables with unit variance, and one finds the standard Hopfield model.

We call the type of disorder generated by (80) a combinatorial disorder. A natural case where it occurs is as follows: Imagine that the patterns are built from a number γN of possible features, where the feature number r is described by the neural activity u_i^r. The variable v_μ^r encodes to what extent feature r belongs to pattern μ. For instance, using binary variables $v_\mu^r = \pm 1$, one can interpret $v_\mu^r = 1$ if and only if feature r belongs to feature μ. Then the pattern μ on site i is (up to an overall constant), by the sum of the features r belonging to μ.

In combinatorial disorder, the random patterns expressed as (80) can be seen as a kind of superposition of features. This is in contrast with usual types of correlations that were studied in previous years, like biased patterns or Gaussian-distributed patterns with a nontrivial correlation matrix. Obviously, the structure of combinatorial disorder can be elaborated further and the features could become themselves combination of subfeatures, and so on.

We shall now develop the mean-field equations for this modified model.

A. Representation with hidden variables

Using the representation (5), the partition function of the modified Hopfield model with combinatorial disorder can be written as

$$Z = \sum_s \int \prod_\mu \frac{d\lambda_\mu e^{-\beta\lambda_\mu^2/2}}{\sqrt{2\pi\beta}} \times \exp\left[\frac{\beta}{\sqrt{\gamma}} \sum_{r=1}^{\gamma N} \left(\frac{\sum_i u_i^r s_i}{\sqrt{N}}\right)\left(\frac{\sum_\mu v_\mu^r \lambda_\mu}{\sqrt{N}}\right)\right]. \tag{81}$$

It is useful to introduce the auxiliary variables

$$U^r = \frac{1}{\sqrt{N}} \sum_i u_i^r s_i \tag{82}$$

and to use the representation

$$1 = \frac{\beta}{2\pi i} \int dU^r d\hat{U}^r \exp\left[\beta \hat{U}^r \left(\frac{1}{\sqrt{N}} \sum_i u_i^r s_i - U^r\right)\right], \tag{83}$$

where the auxiliary variable $\hat{U}^r$ is integrated in the complex plane along the imaginary axis.

Similarly, we introduce the variable

$$V^r = \frac{1}{\sqrt{N}} \sum_\mu v_\mu^r \lambda_\mu \tag{84}$$

and write an integral representation in terms of an auxiliary variable $\hat{V}^r$.

This gives, up to some overall irrelevant constants,

$$\begin{aligned} Z = \sum_s \int \prod_\mu d\lambda_\mu \\ \times \int \prod_r d\vec{t}^r e^{-\frac{\beta}{2}\sum_\mu \lambda_\mu^2 + \beta \sum_{r=1}^{\gamma N} \left(+\frac{U^r V^r}{\sqrt{\gamma}} - \hat{U}^r U^r - \hat{V}^r V^r\right)} \\ \times \exp\left[\frac{\beta}{\sqrt{N}} \sum_{r=1}^{\gamma N} \sum_{i=1}^{N} \hat{U}^r u_i^r s_i + \frac{\beta}{\sqrt{N}} \sum_{r=1}^{\gamma N} \sum_{\mu=1}^{\alpha N} \hat{V}^r v_\mu^r \lambda_\mu \right. \\ \left. + \frac{\beta}{\sqrt{\gamma}} \sum_{r=1}^{\gamma N} U^r V^r\right], \end{aligned} \tag{85}$$

where the variable $\vec{t}^r$ is $\vec{t}^r = (\hat{U}^r, U^r, \hat{V}^r, V^r)$, the integration element is $d\vec{t}^r = d\hat{U}^r dU^r d\hat{V}^r dV^r$, and the integrals overs $\hat{U}^r$ and $\hat{V}^r$ run along the imaginary axis, while those over U^r and V^r are along the real axis.

The representation (85) contains three types of variables:

(i) The N visible neuron variables s_i.

(ii)) The αN pattern variables λ_μ, which are hidden variables.

(iii) The γN "feature variables" $\vec{t}^r$, which build a new layer of hidden variables, interacting with the other two layers.

Figure 4 shows the factor graph for this problem.

B. Belief propagation

Writing the BP equations for the model (85) is a standard (but lengthy) exercise that goes along exactly the same lines as before. It involves eight types of messages running along the edges of the factor graph shown in Fig. 4. These messages are as follows:

$$m_{i\to r}(s_i), \quad \hat{m}_{r\to i}(s_i), \quad m_{\mu\to r}(\lambda_\mu), \quad \hat{m}_{r\to\mu}(\lambda_\mu), \tag{86}$$

$$\hat{m}_{i\to r}(\vec{t}^r), \quad m_{r\to i}(\vec{t}^r), \quad m_{r\to\mu}(\vec{t}^r), \quad \hat{m}_{\mu\to r}(\vec{t}^r). \tag{87}$$

Here and in the following, the letters i,j are indices of the neuron variables running from 1 to N, the letters r,s are indices of the feature variables running from 1 to γN, and the letters μ,ν are indices of the pattern variables running from 1 to αN. Each message is a function of the argument which is written in parenthesis.

We shall not write explicitly the BP equations but proceed directly to the rBP ones, which can be expressed in terms of

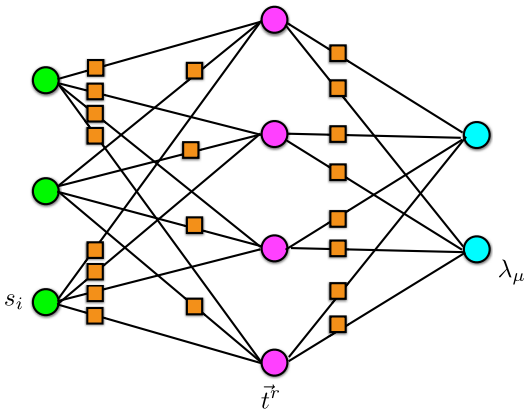

FIG. 4. Factor graph of the Hopfield model with combinatorial patterns in the representation using visible neuron variables (s_i, left layer, green circles), hidden pattern variables (λ_μ, right layer, blue circles), and hidden feature variables (middle layer, purple circles). There exist interaction factors (squares) between each pair of variables belonging to two consecutive layers.

the messages $h_{i\to r}, a_{\mu\to r}, c_{\mu\to r}$ defined from

$$m_{i\to r}(s_i) \cong e^{h_{i\to r} s_i}, \tag{88}$$

$$\int d\lambda_\mu m_{\mu\to r}(\lambda_\mu)\, \lambda_\mu = a_{\mu\to r}, \tag{89}$$

$$\int d\lambda_\mu m_{\mu\to r}(\lambda_\mu)\, \lambda_\mu^2 = c_{\mu\to r} + a_{\mu\to r}^2. \tag{90}$$

They are related through the following set of equations:

$$h_{i\to r} = \frac{1}{\sqrt{N}} \sum_{s(\neq r)} u_i^s f(p_{s\to i}, d_{s\to i}, \pi_s, \delta_s), \tag{91}$$

$$c_{\nu\to r} = \frac{1}{\beta}\left[1 - \frac{1}{N}\sum_{s(\neq r)} \left(v_\nu^s\right)^2 \phi'(p_s, d_s, \pi_{s\to\nu}, \delta_{s\to\nu})\right]^{-1}, \tag{92}$$

$$a_{\nu\to r} = \beta c_{\nu\to r} \frac{1}{N} \sum_{s(\neq r)} v_\nu^s \phi(p_s, d_s, \pi_{s\to\nu}, \delta_{s\to\nu}), \tag{93}$$

where

$$p_r = \frac{1}{\sqrt{N}} \sum_i u_i^r \tanh(\beta h_{i\to r}), \tag{94}$$

$$d_r = \frac{1}{N} \sum_i \left(u_i^r\right)^2 [1 - \tanh^2(\beta h_{i\to r})], \tag{95}$$

$$\pi_r = \frac{1}{\sqrt{N}} \sum_\nu v_\nu^r a_{\nu\to r}, \tag{96}$$

$$\delta_r = \frac{1}{N} \sum_\nu \left(v_\nu^r\right)^2 C_{\nu\to r}, \tag{97}$$

and

$$p_{r\to j} = \frac{1}{\sqrt{N}} \sum_{i(\neq j)} u_i^r \tanh(\beta h_{i\to r}), \tag{98}$$

$$d_{r\to j} = \frac{1}{N} \sum_{i(\neq j)} \left(u_i^r\right)^2 [1 - \tanh^2(\beta h_{i\to r})], \tag{99}$$

$$\pi_{r\to\mu} = \frac{1}{\sqrt{N}} \sum_{\nu(\neq\mu)} v_\nu^r a_{\nu\to r}, \tag{100}$$

$$\delta_{r\to\mu} = \frac{1}{N} \sum_{\nu(\neq\mu)} \left(v_\nu^r\right)^2 c_{\nu\to r}. \tag{101}$$

The functions f, ϕ, ϕ' are functions of four variables defined as

$$f(p,d,\pi,\delta) = \langle \hat{U} \rangle, \tag{102}$$

$$\phi(p,d,\pi,\delta) = \langle \hat{U}^2 \rangle - \langle \hat{U} \rangle^2, \tag{103}$$

$$\phi'(p,d,\pi,\delta) = \frac{\partial}{\partial \pi} \phi(p,d,\pi,\delta), \tag{104}$$

where the expectations are taken with the following measure over $\vec{t} = (U, V, \hat{U}, \hat{V})$:

$$\exp\left[\beta\left(-\hat{U}U - \hat{V}V + \frac{1}{\sqrt{\gamma}} UV + p\hat{U} + \pi\hat{V}\right) + \frac{\beta^2}{2}\left(d\hat{U}^2 + \delta\hat{V}^2\right)\right]. \tag{105}$$

An explicit computation shows that

$$f(p,d,\pi,\delta) = \frac{1}{1 - d\delta\beta^2/\gamma}\left(\frac{\beta\delta}{\gamma} p + \frac{1}{\sqrt{\gamma}}\pi\right), \tag{106}$$

$$\phi(p,d,\pi,\delta) = \frac{1}{1 - d\delta\beta^2/\gamma}\left(\frac{1}{\sqrt{\gamma}} p + \frac{\beta d}{\gamma}\pi\right), \tag{107}$$

$$\phi'(p,d,\pi,\delta) = \frac{1}{1 - d\delta\beta^2/\gamma}\left(\frac{\beta d}{\gamma}\right). \tag{108}$$

C. TAP equations

It turns out that the TAP equations can be written in terms of local quantities associated with each of the variable in the factor graph: Starting from (91)–(93), we define

$$H_i = \frac{1}{\sqrt{N}} \sum_s u_i^s f(p_{s\to i}, d_{s\to i}, \pi_s, \delta_s), \tag{109}$$

$$C_\nu = \frac{1}{\beta}\left[1 - \frac{1}{N}\sum_s \left(v_\nu^s\right)^2 \phi'(p_s, d_s, \pi_{s\to\nu}, \delta_{s\to\nu})\right]^{-1}, \tag{110}$$

$$A_\nu = \beta c_{\nu\to r} \frac{1}{\sqrt{N}} \sum_s v_\nu^s \phi(p_s, d_s, \pi_{s\to\nu}, \delta_{s\to\nu}). \tag{111}$$

We first notice that, in the thermodynamic limit, $C_\nu = C$ becomes independent of ν, and we can also safely approximate $c_{\nu\to r} = C$, the correcting terms being irrelevant. Similarly, we notice that d_r becomes r independent,

$$d_r = 1 - q = 1 - \frac{1}{N} \sum_i \tanh^2(\beta H_i), \tag{112}$$

and δ_r becomes r independent,

$$\delta_r = \alpha C. \tag{113}$$

The equation for C can be obtained from (92),

$$\frac{1}{C} = \beta - \frac{\beta^2(1-q)}{1 - C\alpha\beta^2(1-q)/\gamma}, \tag{114}$$

and gives

$$C = \frac{\gamma}{2\alpha\beta^2(1-q)}\{1-\beta(1-q)(1-\alpha/\gamma) - \sqrt{[1-\beta(1-q)(1+\alpha/\gamma)]^2 - 4\alpha\beta^2(1-q)^2/\gamma}\}, \quad (115)$$

where C is nothing but the variance of the local fields A_ν for each pattern variable. When $\gamma \to \infty$ one finds back that $C = (1/\beta)1/[1-\beta(1-q)]$, which is the expression found in the Hopfield model, as it should.

Defining

$$\hat{f}(p,\pi) = f(p,1-q,\pi,\alpha C) = \frac{1}{1-C\frac{\alpha}{\gamma}\beta^2(1-q)}\left(\frac{\alpha\beta C}{\gamma}p + \frac{1}{\sqrt{\gamma}}\pi\right), \quad (116)$$

$$\hat{\phi}(p,\pi) = \phi(p,1-q,\pi,\alpha C) = \frac{1}{1-C\frac{\alpha}{\gamma}\beta^2(1-q)}\left[\frac{1}{\sqrt{\gamma}}p + \frac{\beta(1-q)}{\gamma}\pi\right], \quad (117)$$

we can write the following TAP equations:

$$H_i = \frac{1}{\sqrt{N}}\sum_s u_i^s\ \hat{f}(p_s,\pi_s) - \frac{\alpha\beta C}{1-C\frac{\alpha}{\gamma}\beta^2(1-q)}\tanh(\beta H_i), \quad (118)$$

$$A_\nu = \frac{1}{N}\sum_s v_\nu^s\ \hat{\phi}(p_s,\pi_s), \quad (119)$$

$$p_r = \frac{1}{\sqrt{N}}\sum_i u_i^r\ \tanh(\beta H_i) - \frac{\beta(1-q)}{\sqrt{\gamma}}\frac{1}{\sqrt{N}}\sum_\nu v_\nu^r\ A_\nu, \quad (120)$$

$$\pi_r = \frac{1}{\sqrt{N}}\sum_\nu v_\nu^r\ A_\nu - \frac{\alpha\beta C}{\sqrt{\gamma}}\frac{1}{\sqrt{N}}\sum_i u_i^r\ \tanh(\beta H_i). \quad (121)$$

Equations (118)–(121), together with the definitions (115)–(117), give the closed set of TAP equations relating the $N(1+\alpha+2\gamma)$ local fields H_i, A_ν, p_r, π_r.

It is interesting to notice that, due to the linear structure of these equations, the variables p_r, π_r can be eliminated explicitly, leading to a set of equations that relate only the fields on the site variables, H_i, and those on the pattern variables, A_μ:

$$H_i = \sum_\nu \frac{\xi_i^\nu}{\sqrt{N}}A_\nu - \frac{\alpha\beta C}{1-C\frac{\alpha}{\gamma}\beta^2(1-q)}\tanh(\beta H_i). \quad (122)$$

$$A_\mu = \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu \tanh(\beta H_j). \quad (123)$$

These TAP equations are similar to the ones of the Hopfield model, with a modified form of the Onsager reaction term. Again, because of their linear structure in A_μ, these variables can be eliminated, giving a set of TAP equation connecting only the local fields on the visible neuron variables:

$$H_i = \frac{1}{N}\sum_j J_{ij}\tanh(\beta H_j) - \frac{\alpha\beta C}{1-C\frac{\alpha}{\gamma}\beta^2(1-q)}\tanh(\beta H_i). \quad (124)$$

Again, the only modification due to the combinatorially correlated patterns is the value of the Onsager reaction term. Notice that, in the large-γ limit, we get back the usual TAP equation of the Hopfield model.

We have derived four versions of the mean-field equations for this modified Hopfield: the rBP equations which relate messages that are propagated on the edges of the factor graph and three versions of the TAP equations: one set of "expanded" equations which relate local quantities associated with each variable node of the factor graph; a second one, "intermediate", which relates the local fields of the neuron variables and the pattern variables; and, finally, the last one that relates only the local fields on the neuron variables. Which one is more useful remains to be investigated. The rBP equations should be studied statistically and give the solution for the thermodynamic properties of this modified Hopfield model using the cavity method. The schedule of update of TAP equations is probably crucial, and working out the correct time indices for algorithmic purpose should go through the expanded version of the equations. It should also be kept in mind that, in general RBMs, the hidden variables are in general not Gaussian distributed, and in such cases the simplification of TAP equations does not occur (see the next section). Therefore, in general, the correct form of TAP equations can be obtained only in their expanded form. This shows the importance of using multilayered networks.

VI. A FEW REMARKS ON MORE GENERAL RESTRICTED BOLTZMANN MACHINES

It is easy to generalize the Hopfield model in order to describe a general RBM. We shall give here the general form of BP, rBP, and TAP equations. Similar results have been obtained recently in information-theoretic approaches to matrix factorization [42,43], but they generally address a form of "planted" problem where specific simplifications take place [13,44]. We give here the general form of the equations.

Using the same notations as before, we consider a system of N neuron variables s_i and P pattern variables λ_μ, described by a probability distribution:

$$P(\{s_i\},\{\lambda_\mu\}) = \frac{1}{Z}\prod_i \tilde{\rho}(s_i)\prod_\mu \rho(\lambda_\mu) \times \exp\left[\beta\left(\sum_i \tilde{h}_i s_i + \sum_\mu h_\mu\lambda_\mu + \sum_{\mu,i}\frac{\xi_i^\mu}{\sqrt{N}}s_i\lambda_\mu\right)\right]. \quad (125)$$

With respect to the usual Hopfield model, three modifications have been introduced:

(i) The local measure on the spin variables is $\tilde{\rho}(s)$. In the Hopfield model one considers $\tilde{\rho}(s) = (1/2)(\delta_{s,1}+\delta_{s,-1})$, but more general distributions can be studied as well.

(ii) The local measure on the pattern variables is $\rho(\lambda)$. In the Hopfield model one considers $\rho(\lambda) = (1/\sqrt{2\pi\beta})\exp(-\beta\lambda^2/2)$, but more general distributions can be studied as well.

(iii) We introduce local fields $\tilde{h}_i$ and h_μ, which will make it possible to compute correlation functions through linear response, using, for instance, $\langle s_i s_j\rangle = \partial\langle s_i\rangle/\partial\tilde{h}_j$.

The BP equations are as follows:

$$m_{i\to\mu}(s_i) \cong \tilde{\rho}(s_i)\, e^{\beta\tilde{h}_i s_i} \prod_{\nu(\neq\mu)} \hat{m}_{\nu\to i}(s_i), \tag{126}$$

$$\hat{m}_{\mu\to i}(s_i) \cong \int d\lambda_\mu m_{\mu\to i}(\lambda_\mu) \exp\left((\beta/\sqrt{N})\xi_i^\mu s_i\lambda_\mu\right), \tag{127}$$

$$\hat{m}_{i\to\mu}(\lambda_\mu) = \int ds_i\; m_{i\to\mu}(s_i) e^{\beta\xi_i^\mu/\sqrt{N}}, \tag{128}$$

$$m_{\mu\to i}(\lambda_\mu) \cong \rho(\lambda_\mu)\, e^{\beta h_\mu\lambda_\mu} \prod_{j(\neq i)} \hat{m}_{j\to\mu}(\lambda_\mu). \tag{129}$$

In order to write the rBP equations, we need to understand the scaling of the variables. In particular, we have seen in the Hopfield model that, in the retrieval phase, one of the variables λ_μ may become very large (of order $\sqrt{N}$), signaling a polarization towards this pattern. The possibility of such a phenomenon clearly depends on the measures $\tilde{\rho}(s)$ and $\rho(\lambda)$. In the Hopfield case, $\rho(\lambda)$ is a Gaussian. This means that the response of the pattern variable λ to a local field h is a linear function of h. This allows the variable λ_μ to grow to very large values. In contrast, in many applications of RBMs, one uses variables with a bounded range of values. For instance, if $\rho(\lambda_\mu)$ vanishes outside an interval $[-C,C]$, then the response of the variable λ_μ is a nonlinear, sigmoid-shaped function of the local field, and the condensation cannot occur.

One opposite case would be the one when both $\tilde{\rho}(s)$ and $\rho(\lambda)$ are Gaussian. It is then clear that, at low temperatures, the spins will acquire spontaneous polarization in the direction of the eigenvector of the J matrix [Eq. (3)] with largest eigenvalue. Both the neuron variables and the pattern variables condense in this case.

We shall write here the rBP equations assuming that there is no condensation. As for the coupling variables ξ_i^μ, we suppose that they are independent identically distributed variables with zero mean and a finite variance.

Following standard procedures like those used in Refs. [14–18], the messages $m_{\mu\to i}(\lambda_\mu)$ and $m_{i\to\mu}(s_i)$ are parameterized in terms of their first two moments. Generalizing (21) and (22), we define

$$a_{\mu\to i} = \int d\lambda_\mu m_{\mu\to i}(\lambda_\mu)\lambda_\mu, \tag{130}$$

$$c_{\mu\to i} = \int d\lambda_\mu m_{\mu\to i}(\lambda_\mu)\lambda_\mu^2 - a_{\mu\to i}^2, \tag{131}$$

$$\tilde{a}_{i\to\mu} = \int ds_i m_{i\to\mu}(s_i) s_i, \tag{132}$$

$$\tilde{c}_{i\to\mu} = \int ds_i m_{i\to\mu}(s_i) s_i^2 - \tilde{a}_{i\to\mu}^2. \tag{133}$$

The rBP equations relating these four types of messages can be written in terms of the following four functions of two real variables.

Considering a neuron variable s with local measure

$$\tilde{P}(s) = \frac{1}{\tilde{z}}\tilde{\rho}(s) e^{us+(v/2)s^2}, \tag{134}$$

we define

$$\tilde{f}(u,v) = \int ds\tilde{P}(s)s, \tag{135}$$

$$\tilde{f}'(u,v) = \frac{\partial}{\partial u}\tilde{f}(u,v) = \int ds\tilde{P}(s)s^2 - \tilde{f}_a(u,v)^2. \tag{136}$$

Considering a pattern variable λ with local measure

$$P(\lambda) = \frac{1}{z}\rho(\lambda) e^{u\lambda+(v/2)\lambda^2}, \tag{137}$$

we define

$$f(u,v) = \int d\lambda P(\lambda)\lambda, \tag{138}$$

$$f'(u,v) = \frac{\partial}{\partial u} f(u,v) = \int d\lambda P(\lambda)\lambda^2 - f_a(u,v)^2. \tag{139}$$

The rBP equations can then be written as

$$a_{\mu\to i}^{t+1} = f\left(\beta h_\mu + \frac{\beta}{\sqrt{N}}\sum_{j(\neq i)} \xi_j^\mu \tilde{a}_{j\to\mu}^t\,,\, \frac{\beta^2}{N}\sum_{j(\neq i)} \left(\xi_j^\mu\right)^2 \tilde{c}_{j\to\mu}^t\right), \tag{140}$$

$$c_{\mu\to i}^{t+1} = f'\left(\beta h_\mu + \frac{\beta}{\sqrt{N}}\sum_{j(\neq i)} \xi_j^\mu \tilde{a}_{j\to\mu}^t\,,\, \frac{\beta^2}{N}\sum_{j(\neq i)} \left(\xi_j^\mu\right)^2 \tilde{c}_{j\to\mu}^t\right), \tag{141}$$

$$\tilde{a}_{i\to\mu}^{t+2} = \tilde{f}\left(\beta\tilde{h}_i + \frac{\beta}{\sqrt{N}}\sum_{\nu(\neq\mu)} \xi_i^\nu a_{\nu\to i}^{t+1}\,,\, \frac{\beta^2}{N}\sum_{\nu(\neq\mu)} \left(\xi_i^\nu\right)^2 c_{\nu\to i}^{t+1}\right), \tag{142}$$

$$\tilde{c}_{i\to\mu}^{t+2} = \tilde{f}'\left(\beta\tilde{h}_i + \frac{\beta}{\sqrt{N}}\sum_{\nu(\neq\mu)} \xi_i^\nu a_{\nu\to i}^{t+1}\,,\, \frac{\beta^2}{N}\sum_{\nu(\neq\mu)} \left(\xi_i^\nu\right)^2 c_{\nu\to i}^{t+1}\right), \tag{143}$$

where we have reintroduced the time indices corresponding to a parallel update of these equations.

One gets the TAP equations using the same method as before. In the large-N limit the messages depend only weakly on the index of arrival. Writing

$$a_{\mu\to i} \simeq A_\mu; \quad c_{\mu\to i} \simeq C_\mu; \quad \tilde{a}_{i\to\mu} \simeq \tilde{A}_i; \quad \tilde{c}_{i\to\mu} \simeq C_i \tag{144}$$

and expanding the leading correction terms, one obtains

$$A_\mu^{t+1} = f\left(U_\mu^t, V_\mu^t\right), \tag{145}$$

$$C_\mu^{t+1} = f'\left(U_\mu^t, V_\mu^t\right), \tag{146}$$

$$\tilde{A}_i^{t+2} = \tilde{f}\left(\tilde{U}_i^{t+1}, \tilde{V}_i^{t+1}\right), \tag{147}$$

$$\tilde{C}_i^{t+2} = \tilde{f}'\left(\tilde{U}_i^{t+1}, \tilde{V}_i^{t+1}\right), \tag{148}$$

where

$$U_\mu^t = \beta h_\mu + \frac{\beta}{\sqrt{N}} \sum_i \xi_i^\mu \tilde{A}_i^t - A_\mu^{t-1} \frac{\beta^2}{N} \sum_i \left(\xi_i^\mu\right)^2 \tilde{f}'\left(\tilde{U}_i^{t-1}, \tilde{V}_i^{t-1}\right), \quad (149)$$

$$V_\mu^t = \frac{\beta^2}{N} \sum_i \left(\xi_i^\mu\right)^2 \tilde{C}_i^t, \quad (150)$$

$$\tilde{U}_i^{t+1} = \beta \tilde{h}_i + \frac{\beta}{\sqrt{N}} \sum_\mu \xi_i^\mu A_\mu^{t+1} - \tilde{A}_i^t \frac{\beta^2}{N} \sum_\mu \left(\xi_i^\mu\right)^2 f'\left(U_\mu^t, V_\mu^t\right), \quad (151)$$

$$\tilde{V}_i^{t+1} = \frac{\beta^2}{N} \sum_\mu \left(\xi_i^\mu\right)^2 C_\mu^{t+1}. \quad (152)$$

Notice that, in general, when $\tilde{\rho}(s)$ and $\rho(\lambda)$ are non-Gaussian, the functions f and $\tilde{f}$ are nonlinear functions of u, and therefore one cannot easily eliminate one of the variables, as was done in the Hopfield model. This means that the correct form of TAP equations require working on the bipartite graph with the two layers of variables, visible and hidden.

VII. CONCLUDING REMARKS

We have seen that the correct mean-field "TAP" equations in the Hopfield model can be written most easily by introducing a layer of hidden variables, the pattern variables, which interact with the neuron variables. In the Hopfield model, the local fields associated with the hidden variables can be eliminated and one remains with TAP equations that are similar to those of general spin glasses, differing only in the detailed form of the Onsager reaction term. However, when one deals with RBMs which generalize the Hopfield model with non-Gaussian hidden variables, the representation with the hidden layer is necessary.

In the case where the patterns to be memorized have correlations based on a combinatorial structure, the TAP equations involve one extra layer of hidden variables, and with a deeper structure of correlations, extra hidden layers would be added.

We believe that combinatorial disorder is actually an essential ingredient that is likely to be present in real data. In this respect, it is striking that the correct treatment of mean-field theory in RBMs with combinatorial disorder leads naturally to the appearance of layers of hidden variables. The present study of the Hopfield model is a kind of first test of this idea, which we hope could lead to a better understanding of the role of multilayered structures in practical applications of neural networks.

The present work calls for some further developments in several directions:

(i) It will be interesting to study how the TAP estimates for the magnetizations (and those for the correlation functions that are inferred through linear response) can be turned into efficient algorithms for unsupervised learning, along the lines of Refs. [6–10]. In this respect, it is interesting to be able to study controlled problems. We think that the Hopfield model with combinatorial-correlated patterns can be used as an interesting teacher to generate data, i.e., patterns of neural activity, that can be used in the training of a "student" Hopfield network.

(ii) The modified Hopfield model with combinatorial-correlated patterns is interesting in itself. It would be interesting to study its thermodynamics both with replicas and with the cavity method through a statistical analysis of the rBP equations.

ACKNOWLEDGMENTS

It is a pleasure to thank C. Baldassi, F. Krzakala, L. Zdeborová, and R. Zecchina for interesting discussions related to the subject of this paper.

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**, 436 (2015).

[2] R. Salakhutdinov, A. Mnih, and G. Hinton, Restricted boltzmann machines for collaborative filtering, *Proceedings of the 24th International Conference on Machine Learning* (ACM, New York, 2007), pp. 791–798.

[3] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, *Proceedings of the 25th International Conference on Machine Learning* (ACM, New York, 2008), pp. 1096–1103.

[4] M. Welling and G. E. Hinton, A new learning algorithm for mean field boltzmann machines, *International Conference on Artificial Neural Networks* (Springer, Berlin, 2002), pp. 351–357.

[5] T. Tieleman, Training restricted boltzmann machines using approximations to the likelihood gradient, *Proceedings of the 25th International Conference on Machine Learning* (ACM, New York, 2008), pp. 1064–1071.

[6] H. J. Kappen and F. Rodriguez, Boltzmann machine learning using mean field theory and linear response correction, *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 1998), pp. 280–286.

[7] T. Tanaka, Mean-field theory of boltzmann machine learning, Phys. Rev. E **58**, 2302 (1998).

[8] H. Huang and T. Toyoizumi, Advanced mean-field theory of the restricted boltzmann machine, Phys. Rev. E **91**, 050101 (2015).

[9] M. Gabrié, E. W. Tramel, and F. Krzakala, Training restricted boltzmann machine via the thouless-anderson-palmer free energy, *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2015), pp. 640–648.

[10] E. W. Tramel, A. Manoel, F. Caltagirone, M. Gabrié, and F. Krzakala, Inferring sparsity: Compressed sensing using

generalized restricted boltzmann machines, *Information Theory Workshop (ITW)* (IEEE, 2016).
[11] T. Richardson and R. Urbanke, *Modern Coding Theory* (Cambridge University Press, Cambridge, 2008).
[12] M. Mézard and A. Montanari, *Information, Physics and Computation* (Oxford University Press, Oxford, 2009).
[13] L. Zdeborová and F. Krzakala, Statistical physics of inference: Thresholds and algorithms, Adv. Phys. **65**, 453 (2016).
[14] M. Bayati and A. Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing, IEEE Trans. Inf. Theory **57**, 764 (2011).
[15] S. Rangan, Estimation with random linear mixing, belief propagation and compressed sensing, *44th Annual Conference on Information Sciences and Systems (CISS)* (IEEE, Los Alamitos, CA, 2010), pp. 1–6.
[16] S. Rangan, Generalized approximate message passing for estimation with random linear mixing, *IEEE International Symposium on Information Theory Proceedings (ISIT)* (IEEE, Los Alamitos, CA, 2011), pp. 2168–2172.
[17] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová, Statistical-Physics-Based Reconstruction in Compressed Sensing, Phys. Rev. X **2**, 021005 (2012).
[18] 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.
[19] M. Mézard, The space of interactions in neural networks: Gardner's computation with the cavity method, J. Phys. A **22**, 2181 (1989).
[20] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina, Efficient supervised learning in networks with binary synapses, BMC Neurosci. **8**, S13 (2007).
[21] C. Baldassi and A. Braunstein, A max-sum algorithm for training discrete neural networks, J. Stat. Mech.: Theor. Exp. (2015) P08008.
[22] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).
[23] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[24] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Solution of 'solvable model of a spin glass', Philos. Mag. **35**, 593 (1977).
[25] Y. Kabashima and D. Saad, The tap approach to intensive and extensive connectivity systems, Advanced Mean Field Methods-Theory and Practice **6**, 65 (2001).
[26] K. Nakanishi and H. Takayama, Mean-field theory for a spin-glass model of neural networks: Tap free energy and the paramagnetic to spin-glass transition, J. Phys. A **30**, 8085 (1997).
[27] M. Shamir and H. Sompolinsky, Thouless-anderson-palmer equations for neural networks, Phys. Rev. E **61**, 1839 (2000).
[28] D. Sherrington and S. Kirkpatrick, Solvable Model of a Spin-Glass, Phys. Rev. Lett. **35**, 1792 (1975).
[29] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Spin-glass models of neural networks, Phys. Rev. A **32**, 1007 (1985).
[30] 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).
[31] Y. Kabashima, A cdma multiuser detection algorithm on the basis of belief propagation, J. Phys. A **36**, 11111 (2003).
[32] T. Tanaka and M. Okada, Approximate belief propagation, density evolution, and statistical neurodynamics for cdma multiuser detection, IEEE Trans. Inf. Theory **51**, 700 (2005).
[33] A. Montanari and D. Tse, Analysis of belief propagation for non-linear problems: The example of cdma (or: How to prove tanaka's formula), *2006 IEEE Information Theory Workshop-ITW'06 Punta del Este* (IEEE, Los Alamitos, CA, 2006), pp. 160–164.
[34] D. Guo and C.-C. Wang, Asymptotic mean-square optimality of belief propagation for sparse linear systems, *Information Theory Workshop, 2006* (ITW '06) (IEEE, Los Alamitos, CA, 2006), pp. 194–198.
[35] D. L. Donoho, A. Maleki, and A. Montanari, Message-passing algorithms for compressed sensing, Proc. Natl. Acad. Sci. USA **106**, 18914 (2009).
[36] J. T. Parker, V. Cevher, and P. Schniter, Compressive sensing under matrix uncertainties: An approximate message passing approach, *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)* (IEEE, Los Alamitos, CA, 2011), pp. 804–808.
[37] P. Schniter, J. Parker, and V. Cevher, Bilinear generalized approximate message passing (big-amp) for matrix recovery problem, *Workshop on Information Theory and Applications (ITA)* (IEEE, Los Alamitos, CA, 2012).
[38] V. A. Marcenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR Sb. **1**, 457 (1967).
[39] E. Bolthausen, An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model, Commun. Math. Phys. **325**, 333 (2014).
[40] A. Braunstein and R. Zecchina, Learning by Message Passing in Networks of Discrete Synapses, Phys. Rev. Lett. **96**, 030201 (2006).
[41] M. Opper and O. Winther, Adaptive and self-averaging thouless-anderson-palmer mean-field theory for probabilistic modeling, Phys. Rev. E **64**, 056131 (2001).
[42] Y. Deshpande and A. Montanari, Sparse pca via covariance thresholding, *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2014), pp. 334–342.
[43] T. Lesieur, F. Krzakala, and L. Zdeborov, Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel, *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE, Los Alamitos, CA, 2015), pp. 680–687.
[44] F. Krzakala and L. Zdeborová, Hiding Quiet Solutions in Random Constraint Satisfaction Problems, Phys. Rev. Lett. **102**, 238701 (2009).