Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling

Opper, Manfred; Winther, Ole

Published in:
Physical Review E. Statistical, Nonlinear, and Soft Matter Physics

Link to article, DOI:
10.1103/PhysRevE.64.056131

Publication date:
2001

Document Version
Publisher's PDF, also known as Version of record

Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling

Manfred Opper[1] and Ole Winther[2]
[1]*Neural Computing Research Group, School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, United Kingdom*
[2]*Center for Biological Sequence Analysis, BioCentrum DTU, Technical University of Denmark, B208, 2800 Lyngby, Denmark; Informatics and Mathematical Modelling, Technical University of Denmark, B321, 2800 Lyngby, Denmark; and Theoretical Physics, Lund University, Sölvegatan 14 A, 223 62 Lund, Sweden*
(Received 10 May 2001; published 30 October 2001)

We develop a generalization of the Thouless-Anderson-Palmer (TAP) mean-field approach of disorder physics, which makes the method applicable to the computation of approximate averages in probabilistic models for real data. In contrast to the conventional TAP approach, where the knowledge of the distribution of couplings between the random variables is required, our method adapts to the concrete set of couplings. We show the significance of the approach in two ways: Our approach reproduces replica symmetric results for a wide class of toy models (assuming a nonglassy phase) with given disorder distributions in the thermodynamic limit. On the other hand, simulations on a real data model demonstrate that the method achieves more accurate predictions as compared to conventional TAP approaches.

DOI: 10.1103/PhysRevE.64.056131 PACS number(s): 02.50.−r

I. INTRODUCTION

Models of statistical physics with random infinite ranged interactions have the nice feature that they can be treated exactly by mean-field methods. To compute average properties of the system, one may choose two possible alternative but equivalent approaches. The first one is based on the replica method in which the replicated system is first averaged over the disorder and the resulting nonrandom system is decoupled by saddle-point methods, which leads to exact average case mean-field equations [1]. In the second approach one derives a mean-field theory for a *fixed set of random couplings*, which becomes exact in the thermodynamic limit for almost all realizations of the randomness. This type of mean-field theory is traditionally called the TAP approach after Thouless, Anderson, and Palmer [2] who developed it first for the Sherrington-Kirkpatrick model [3] of spin glasses. In a final step, one may average TAP mean-field equations over the couplings to achieve the same result as in the replica approach [4].

Besides the importance of the TAP approach in the theory of disordered systems there is a recent interest in this method, which comes from a more applied area of research dealing with adaptive probabilistic data models (for a review see, e.g., [5]). The goal of such models is to explain complex observed data by a set of unobserved, *hidden* random variables based on the joint distribution of both sets of variables. A few popular examples are Bayes belief networks [6] (used as trainable expert systems), independent component analysis [7] (abbreviated ICA, which detects independent sources in nonlinear signal processing), Gaussian process models [8] (modeling hidden spatial structures by random fields), and Boltzmann machines [9] (the Ising version of the random fields).

The price that a modeler has to pay for the high degree of flexibility of these models is the vast increase in computational complexity when the number of hidden variables is large. Both the statistical inference about the hidden variables and the learning of the model parameters requires the computation of marginal distributions of the hidden variables/the observed data, i.e., the evaluation of high-dimensional sums or integrals. Since similar types of calculations are ubiquitous in computing thermal averages, e.g., for finding local magnetizations and free energies, there is a great deal of interest in adopting approximation techniques from statistical physics. Already the simple (often called naive) mean-field (MF) method, which neglects all correlations of random variables has been applied successfully to a variety of probabilistic data models. At present, there is a growing research activity in the field of probabilistic models trying to overcome the limitations of the simple MF method by partly including the dependencies of variables but still keeping the approximation tractable. In the case where the individual dependencies are weak but their total effect cannot be neglected, the TAP method is a natural candidate for such an improved approximation. TAP approaches for different probabilistic models have already been discussed for neural networks [10–12], Boltzmann machines [13,14], Gaussian process models for classification [8], error correcting codes [15], etc.; for a review see also [16].

Unfortunately, the TAP mean-field approach shows a characteristic difference from the naive MF method, which makes its straightforward application to models for real data nontrivial. While the simple MF equations are expressed in terms of the concrete couplings (which encode observed data in applications), the *Onsager correction* to the naive MF theory (for models with extensive connectivities) provided by the TAP approach will explicitly depend on the distribution from which these couplings were generated at random. Two models with the same connectivities but different distributions for the couplings, like, e.g., the SK model and the Hopfield model [17] have different expressions for the Onsager corrections (see, e.g., [1], Chap. XIII). While in the

models of disorder physics this distribution is given in advance, such a knowledge is obviously not available for models of real data. Simply taking results from a theory that *assumes* a specific distribution may lead to suboptimal performance. To define the ''correct'' TAP approach, which is valid in these more general situations, truncated perturbative expansions for the free energy or for marginal distributions have been considered [18–21]. While such a finite order truncation becomes exact for the SK model this is, in general, not to be expected.

In this paper we present a new approach to this problem.[1] Our criteria for a valid TAP method are twofold: We require that the lack of knowledge of the underlying distribution of the couplings must be compensated by a self-consistent computation, which *adapts* the Onsager correction to the *concrete* set of couplings. Second, when applied to a set of interactions, which was randomly generated from a *known* distribution (for which the mean-field assumption is valid), a suitable average of the adaptive TAP method should reproduce the *correct* average case results known from the replica approach. We achieve the first goal by combining the cavity approach [1] with a simple linear response technique, which yields a second set of TAP equations for the Onsager correction. This method fulfills the second requirement so far only for the class of extensively connected models with nonglassy behavior. These models can be described by a single ergodic phase, which is correctly described by a finite set of order parameters (unlike the sparsely connected models [1]) in replica symmetry. Our experience with average case studies of neural networks makes us expect that the assumption of replica symmetry may well describe practical situations when the models are sufficiently matched to the data.

Our approach is most naturally developed for models with pairwise interactions between variables S_i, $i=1, \ldots, N$

$$P(\mathbf{S})=\frac{\rho(\mathbf{S})}{Z(\boldsymbol{\theta},\mathbf{J})}\exp\left[\sum_{i<j} S_i J_{ij} S_j+\sum_i S_i\theta_i\right]. \tag{1}$$

Here $\mathbf{S}=(S_1,\ldots,S)$, and we have set $J_{ii}=0$. All self-interactions are contained in the factorizing distribution $\rho(\mathbf{S})\equiv\Pi_j\rho_j(S_j)$, which also contains all single variable constraints of the variables S_i like their range, their discreteness, etc. Examples of models that are included in this framework are Ising models (like the SK model, the Hopfield model, the Boltzmann machine in the neural computation context), the finite temperature versions of the matching and traveling salesman problems [23,1], Gaussian process models [8], and the ICA model of [24]. However, many other interesting data models are of a more complicated form such as

$$P(\mathbf{S})\propto\rho(\mathbf{S})\exp\left[\sum_{i<j} S_i J_{ij} S_j+\sum_i S_i\theta_i\right]\prod_{k=1}^{m} F\left(\sum_{i=1}^{N}\hat{J}_{ki}S_i\right), \tag{2}$$

which includes a variety of popular ''network'' models like perceptrons (see Sec. IV), sigmoid belief networks [25,22,26], and combinatorial optimization problems with inequality constraints, e.g., the knapsack problem [27].

Luckily, the models of the type Eq. (2) can be easily represented in the standard form Eq. (1) by the ''field theoretic'' trick of introducing the fields $\Sigma_{i=1}^{N}\hat{J}_{ki}S_i$, $k=1, \ldots, m$ as new variables by using Dirac δ functions and their exponential representations. Denoting the purely imaginary conjugate variables by $\hat{\mathbf{S}}=(\hat{S}_1,\ldots,\hat{S}_m)$, the space of variables is augmented to the set $(\mathbf{S},\hat{\mathbf{S}})$ where the ''prior distribution'' for the hatted variables is given by

$$\hat{\rho}(\hat{S})=\int\frac{d\hat{h}}{2\pi i}e^{-\hat{S}\hat{h}}F(\hat{h}) \tag{3}$$

and the augmented coupling matrix is

$$\mathbf{J}_{\mathrm{aug}}=\begin{pmatrix}\mathbf{J} & \hat{\mathbf{J}}^T\\ \hat{\mathbf{J}} & 0\end{pmatrix}. \tag{4}$$

Since all the subsequent manipulations are of the analytic type, i.e., they are based on certain formal expansions rather than on probabilistic arguments (in the sense of assuming positive normalized measures), we expect that our use of nonpositive and even complex measures will not be problematic.

The paper is organized as follows. The adaptive TAP equations are derived in Sec. II (with a summary given in Sec. II E). In Sec. III, we show how the adaptive theory reproduces the correct average case results when applied to a fairly general class of distributions for the couplings. We derive ''self-averaging'' TAP equations, replica results, and the stability condition for the mean-field solution [the de Almeida–Thouless (AT) condition]. In Sec. IV, we apply our results to the SK model, the Hopfield network, and the simple perceptron. Finally, we present an outlook in Sec. V.

II. ADAPTIVE TAP APPROACH

In this section, we will derive both an adaptive TAP approximation for the marginal distribution $P_i(S)\equiv\int\Pi_{j\neq i}dS_jP(\mathbf{S})$ (Secs. II A–II C) and the free energy $F(\mathbf{J},\boldsymbol{\theta})=-\ln Z(\mathbf{J},\boldsymbol{\theta})$ (Sec. II D). The free energy corresponds to the negative log probability of the observed data, which can be used as a yardstick for deciding which model best fits the data.

Our derivation will be based on the cavity approach introduced by [1]. We will assume that we are not dealing with a glassy system with its many ergodic components, but that all averages are for a single state. This is usually expected to hold when the probabilistic model is well matched to the data. We expect that the adaptive TAP approximation can be extended to glassy systems along the line of Chap. V in [1].

A. The marginal distribution

The starting point of our derivation is the following exact equation for the marginal distribution of the variable S_i

[1]A shorter presentation of the main results of this paper can be found in [22].

$$P_i(S_i)=\frac{\int \prod_{j,j\neq i} dS_j \rho_i(S_i)\exp\left[S_i\left(\sum_j J_{ij}s_j+\theta_i\right)\right]p(\mathbf{S}\backslash S_i)}{\int \prod_j dS_j \rho_i(S_i)\exp\left[S_i\left(\sum_j J_{ij}S_j+\theta_i\right)\right]P(\mathbf{S}\backslash S_i)}, \tag{5}$$

where $P(\mathbf{S}\backslash S_i)$ is the distribution of all variables for a system where the ith variable is absent. We see from Eq. (5) that S_i interacts with the remaining variables only through the field $h_i=\Sigma_j J_{ij}S_j$. Hence, we introduce its ''cavity'' distribution, i.e., the distribution of the field at the ''position'' of the ''empty'' site i by

$$P(h_i\backslash S_i)=\int \prod_{j\neq i} dS_j\, \delta\left(h_i-\sum_j J_{ij}S_j\right)P(\mathbf{S}\backslash S_i) \tag{6}$$

and rewrite Eq. (5) as

$$P_i(S_i)=\frac{\rho_i(S_i)}{Z_0^{(i)}}e^{-H_i(S_i)}, \tag{7}$$

where we have introduced an effective single variable Hamiltonian

$$-H_i(S)=\ln\langle e^{Sh_i}\rangle_{\backslash i}, \tag{8}$$

and the brackets $\langle\cdots\rangle_{\backslash i}$ denote an average with respect to $P(h_i\backslash S_i)$, Eq. (6). The corresponding partition function is

$$Z_0^{(i)}=\int dS\, \rho_i(S)e^{-H_i(S)}, \tag{9}$$

The complete knowledge of $H_i(S)$ would provide us with the ability to compute averages of functions of a single variable S_i like, e.g.,

$$\langle S_i\rangle=\frac{\partial}{\partial\theta_i}\ln Z_0^{(i)}. \tag{10}$$

Second, by using appropriate derivatives with respect to the external fields θ_j we can compute correlation functions. For example, the connected correlation function of two variables is expressed by the *linear response relation* as

$$\chi_{ij}\equiv\langle S_iS_j\rangle-\langle S_i\rangle\langle S_j\rangle=\frac{\partial\langle S_i\rangle}{\partial\theta_j}=\frac{\partial^2\ln Z_0^{(i)}}{\partial\theta_i\partial\theta_j}. \tag{11}$$

Moreover, from the definition (8), we also realize that the distribution of h_i can be reconstructed using derivatives with respect to S. A corresponding result that will be useful in the following is

$$\langle h_i\rangle=\frac{1}{Z_0^{(i)}}\int dS\, \rho_i(S)\frac{\partial}{\partial S}e^{-H_i(S)}. \tag{12}$$

B. The cavity approach

In general, we can express Eq. (8) by the cumulants $\kappa_k^{(i)}$ of the *cavity* distribution

$$-H_i(S)=\sum_{k=1}^{\infty}\frac{\kappa_k^{(i)}}{k!}S^k \tag{13}$$

with $\kappa_1^{(i)}=\langle h_i\rangle_{\backslash i}$, $\kappa_2^{(i)}=\langle h_i^2\rangle_{\backslash i}-\langle h_i\rangle_{\backslash i}^2$, etc.

The usual argument in deriving TAP equations is based on the assumption of weak dependencies between the random variables $\mathbf{S}$, which is expressed in the so-called clustering hypothesis [1],

$$\frac{1}{N^2}\sum_{ij}\left(\langle S_iS_j\rangle-\langle S_i\rangle\langle S_j\rangle\right)^2\to 0 \tag{14}$$

and similar relations for higher connected correlation functions. One concludes that if the variable S_i is removed from the system, the effect of the correlations between the $S_j's$ on the distribution of the field h_i is so weak that (as in proofs of the central limit theorem by characteristic functions) one can neglect all cumulants of order greater than 2, i.e.,

$$\kappa_k^{(i)}=0 \quad \text{for } k>2. \tag{15}$$

If $\mathbf{S}$ is a real random vector this is equivalent to approximating Eq. (6) by a Gaussian distribution [1,8], setting

$$P(h_i\backslash S_i)\approx\frac{1}{\sqrt{2\pi V_i}}\exp\left[-\frac{(h_i-\langle h_i\rangle_{\backslash i})^2}{2V_i}\right], \tag{16}$$

where $V_i\equiv\kappa_2^{(i)}=\langle h_i^2\rangle_{\backslash i}-\langle h_i\rangle_{\backslash i}^2$. From Eq. (15), we immediately get the marginal distribution

$$P_i(S)=\frac{1}{Z_0^{(i)}}\rho_i(S)\exp\left[S(\langle h_i\rangle_{\backslash i}+\theta_i)+\frac{V_i}{2}S^2\right] \tag{17}$$

and the single variable partition function

$$Z_0^{(i)}=\int dS\, \rho_i(S)\exp\left[S(\langle h_i\rangle_{\backslash i}+\theta_i)+\frac{V_i}{2}S^2\right]. \tag{18}$$

In the following, we will assume the validity of Eq. (15) and the resulting Eq. (17) also in the cases, where S is a complex variable with nonreal prior distribution $\rho(S)$.

All that remains to derive the TAP equations is to compute the sets of the first two cumulants $\langle h_i\rangle_{\backslash i}$ and V_i for $i=1,\ldots,N$ self-consistently. The first cumulant is easily found from Eq. (12) using Eq. (17) as

$$\langle h_i\rangle=\langle h_i\rangle_{\backslash i}+V_i\langle S_i\rangle. \tag{19}$$

Hence, we can eliminate $\langle h_i\rangle_{\backslash i}$ in favor of the mean-field variables $\langle S_i\rangle$ and $V_i, i=1,\ldots,N$ via

$$\langle h_i\rangle_{\backslash i}=\sum_j J_{ij}\langle S_j\rangle-V_i\langle S_i\rangle. \tag{20}$$

This has the well-known structure of a mean field corrected by a so-called *Onsager reaction* term, which accounts for the nontrivial correlations between variables that are neglected in a naive mean-field approach.

So far, this approach is well known. The new aspect of our paper is in the way we compute the V_i's. A naive computation would lead to

$$V_i = \sum_{j,k} J_{ij}J_{ik}(\langle S_j S_k\rangle_{\setminus i} - \langle S_j\rangle_{\setminus i}\langle S_k\rangle_{\setminus i}) \approx \sum_j J_{ij}^2(\langle S_j^2\rangle_{\setminus i} - \langle S_j\rangle_{\setminus i}^2), \tag{21}$$

neglecting the nondiagonal correlations. This turns out to be correct in the thermodynamic limit $N\to\infty$ for models (like the SK model) where different pairs of couplings J_{ij} and J_{ik} are drawn *independently* at random. However, it fails when the J_{ij}'s become weakly correlated. Consider, e.g., the simple Gaussian model $\rho(\mathbf{S})\propto e^{-(1/2)|\mathbf{S}|^2}$ with a ''Hopfield-type'' coupling matrix defined by $J_{ij}=(1/N)\ \Sigma_{k=1}^{\alpha N} x_i^k x_j^k$, where x_i^k are random variables with zero mean and unit variance. The corresponding covariance

$$\langle S_i S_j\rangle - \langle S_i\rangle\langle S_j\rangle = (\mathbf{I}-\mathbf{J})_{ij}^{-1} \tag{22}$$

would still fulfill the clustering condition (14), but nondiagonal contributions to V_i do not vanish because of higher order correlations, e.g., $\Sigma_{jk}J_{ij}J_{ik}J_{jk}\approx\alpha$ do not vanish.

While it is well known how to derive TAP equations for this type of coupling matrix $\mathbf{J}$ (see, e.g., [1,28]), these approaches require the explicit knowledge of the statistics of the J_{ij}'s. In the following section, we aim at a computation of the cavity field variances V_i's, which does not assume such a knowledge.

C. Computing the second moments

Our derivation is based on a self-consistent computation of the matrix of susceptibilities $\chi_{ij}=\partial\langle S_i\rangle/\partial\theta_j=\langle S_iS_j\rangle - \langle S_i\rangle\langle S_j\rangle$ based on the mean Eq. (10). Hence, the diagonal elements χ_{ii}, $i=1,\ \ldots,\ N$ are expressed both by linear response and by the explicit result $\chi_{ii}=\langle S_i^2\rangle-\langle S_i\rangle^2$ that can be evaluated using the marginal distribution (17). Equating the two expressions we obtain implicit equations for the variances $V_i=\langle h_i^2\rangle_{\setminus i}-\langle h_i\rangle_{\setminus i}^2$, $i=1,\ \ldots,\ N$. Self-interactions $V_i\langle S_i\rangle$ determined by the linear response method have also been introduced in [29] as a heuristics to correct the naive MF equations for Boltzmann machines.

Our crucial approximation in the linear response calculation is that it is sufficient to include a perturbation of the means of the cavity fields $\langle h_i\rangle_{\setminus i}$ whereas the variances V_i are kept unchanged. This is consistent with the fact that the V_i's become self-averaging quantities in a suitable thermodynamic limit framework where the mean-field method becomes exact. Hence, by differentiating Eq. (10) with respect to the external field θ_j, Eq. (18), we get

$$\chi_{ij}=\frac{\partial\langle S_i\rangle}{\partial\theta_i}\left(\delta_{ij}+\frac{\partial\langle h_i\rangle_{\setminus i}}{\partial\theta_j}\right),$$

where $\partial\langle S_i\rangle/\partial\theta_i$ is the explicit derivative of Eq. (10). Further differentiating Eq. (20), we finally get

$$\chi_{ij}=\frac{\partial\langle S_i\rangle}{\partial\theta_i}\left[\delta_{ij}+\sum_k\ (J_{ik}-V_k\delta_{ik})\chi_{kj}\right], \tag{23}$$

which can be solved with respect to $\boldsymbol{\chi}=\{\chi_{ij}\}$ and yields

$$\boldsymbol{\chi}=(\boldsymbol{\Lambda}-\mathbf{J})^{-1}. \tag{24}$$

Here we have introduced the diagonal matrix

$$\boldsymbol{\Lambda}=\mathrm{diag}(\Lambda_1,\ldots,\Lambda_N),\quad \Lambda_i\equiv V_i+1/\chi_{ii}. \tag{25}$$

Specializing to the diagonal elements $\chi_{ii}=\langle S_i^2\rangle-\langle S_i\rangle^2$ determines V_i implicitly via

$$\chi_{ii}=[(\boldsymbol{\Lambda}-\mathbf{J})^{-1}]_{ii}. \tag{26}$$

The requirement that the susceptibility matrix (i.e., the matrix of covariances) must be positive definite can be used to test whether the mean-field solution is consistent. In the thermodynamic limit this requirement leads to a criterion that is equivalent to the well-known *de Almeida–Thouless* stability condition [1] (see Sec. III D).

Two important remarks should be made at this point: Although all approximations are expected to be exact, in general, only in a suitable thermodynamic limit framework (see Sec. III), the final result (26) is correct for a Gaussian model for *arbitrary N* (see Appendix A). Secondly, we note that the functional relationship between the χ_{ii} and the V_i, Eq. (26), is independent of the specific single spin measure $\rho(S)$. This argument can be used for a derivation of the free energy different from the one of the following sections.

D. The adaptive TAP free energy

In this section, we will derive a TAP approximation to the free energy

$$F(\mathbf{J},\boldsymbol{\theta})=-\ln Z(\mathbf{J},\boldsymbol{\theta})$$

using the adaptive form of the Onsager term given by Eq. (26). For this purpose, it is useful to generalize the model Eq. (1) to a one parameter class of models where the interaction $\mathbf{J}$ is replaced by $l\mathbf{J}$ with $0\leqslant l\leqslant 1$, i.e.,

$$Z(l\mathbf{J},\boldsymbol{\theta})=\int d\mathbf{S}\prod_i\ \rho_i(S_i)\exp\left(\frac{l}{2}\mathbf{S}^T\mathbf{J}\mathbf{S}+\mathbf{S}^T\boldsymbol{\theta}\right). \tag{27}$$

Since the solutions of the TAP equations provide us with the moments $m_i\equiv\langle S_i\rangle$ and $M_i\equiv\langle S_i^2\rangle$ for $i=1,\ \ldots,\ N$, we will work with the *Gibbs free energy*, i.e., the free energy for *fixed* m_i and M_i, which is defined by a Legendre transform using external fields γ_i and λ_i conjugate to m_i and M_i, i.e.,

$$\Phi_l(\mathbf{m},\mathbf{M})=\operatorname*{extr}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}\Psi_l(\boldsymbol{\lambda},\boldsymbol{\gamma},\mathbf{m},\mathbf{M}), \tag{28}$$

where

$$\Psi_l(\boldsymbol{\lambda},\boldsymbol{\gamma},\mathbf{m},\mathbf{M}) = -\ln Z(l\mathbf{J}+\boldsymbol{\lambda},\boldsymbol{\theta}+\boldsymbol{\gamma}) + \sum_i \gamma_i m_i + \sum_i \frac{\lambda_i}{2} M_i . \tag{29}$$

$\boldsymbol{\lambda}$ is a diagonal matrix with entries λ_i. The values $\mathbf{m}^e$ and $\mathbf{M}^e$, which make Φ_l stationary, i.e., for which $\partial_{\mathbf{m}}\Phi_l = \partial_{\boldsymbol{M}}\Phi_l = 0$, determine the correct equilibrium expectation values: $\langle S_i\rangle_l = m_i^e$ and $\langle S_i^2\rangle_l = M_i^e$ (where the index indicates that the expectation is taken with parameter l). Furthermore, by inserting these values back into Φ_1, the original free energy $F(\mathbf{J},\boldsymbol{\theta}) = -\ln Z(\mathbf{J},\boldsymbol{\theta}) = \Phi_1(\mathbf{m}^e,\mathbf{M}^e)$ is recovered.

We compute the TAP approximation to Φ_1 from the relation

$$\Phi_1 = \Phi_0 + \int_0^1 dl \frac{\partial \Phi_l}{\partial l} = \Phi_0 - \frac{1}{2}\int_0^1 dl \left\{ \sum_{i,j} m_i J_{ij} m_j + \mathrm{Tr}(\boldsymbol{\chi}_l \mathbf{J}) \right\} \tag{30}$$

with $\chi_{l,ij} = \langle S_i S_j\rangle_l - \langle S_i\rangle_l \langle S_j\rangle_l$ and

$$\Phi_0 = \operatorname*{extr}_{\boldsymbol{\lambda}^0,\boldsymbol{\gamma}^0} \left\{ -\ln Z(\boldsymbol{\lambda}^0,\boldsymbol{\theta}+\boldsymbol{\gamma}^0) + \sum_i m_i \gamma_i^0 + \sum_i \frac{\lambda_i^0}{2} M_i \right\}. \tag{31}$$

Note, that the derivatives of λ_i and γ_i with respect to l disappear from Eq. (30) because of $\partial_{\gamma_i}\Psi = \partial_{\lambda_i}\Psi = 0$.

We next insert our TAP approximation $\boldsymbol{\chi}_l = (\boldsymbol{\Lambda}_l - l\mathbf{J})^{-1}$, Eq. (24) with $\Lambda_{l,i} = V_{l,i} + 1/\chi_{ii}$, into Eq. (30) and integrate with respect to l. Note, that $\chi_{ii} = M_i - m_i^2$ is a fixed quantity that does not depend on l. Using

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{\chi}_l\mathbf{J}) &= \mathrm{Tr}[(\boldsymbol{\Lambda}_l - l\mathbf{J})^{-1}\mathbf{J}] \\ &= -\frac{d}{dl}\mathrm{Tr}\ln(\boldsymbol{\Lambda}_l - l\mathbf{J}) + \mathrm{Tr}\left(\boldsymbol{\chi}_l \frac{\partial \boldsymbol{\Lambda}_l}{\partial l}\right) \\ &= -\frac{d}{dl}\mathrm{Tr}\ln(\boldsymbol{\Lambda}_l - l\mathbf{J}) + \sum_i \chi_{ii}\frac{\partial \Lambda_{l,i}}{\partial l},\end{aligned}$$

and noting that $V_{0,i} = 0$, we obtain

$$\Phi_1 = \Phi_0 - \tfrac{1}{2}\sum_{ij} m_i J_{ij} m_j + \Delta\Phi, \tag{32}$$

$$\Delta\Phi = \tfrac{1}{2}\ln\det(\boldsymbol{\Lambda} - \mathbf{J}) - \tfrac{1}{2}\sum_i V_i\chi_{ii} + \tfrac{1}{2}\sum_i \ln\chi_{ii}.$$

The first two terms constitute the naive mean-field approximation to Φ and the last term $\Delta\Phi$ is the Onsager correction. Note, that this result is not equivalent to truncation of a power series expansion of Φ to second order in l (often termed the Plefka expansion [18]) but contains terms of all orders. It is easy to see that we also recover the TAP equations from the equilibrium conditions $\partial_{\mathbf{m}}\Phi_1 = \partial_{\boldsymbol{M}}\Phi_1 = 0$. We finally identify the conjugate fields as the mean and the variance of the cavity field via $\gamma_i^0 = \langle h_i\rangle_{\setminus i}$ and $\lambda_i^0 = V_i$ from Eq. (31).

In Sec. III we give a simplification of the free energy for the thermodynamic limit when the distribution of the couplings $\mathbf{J}$ is explicitly given, i.e., for the conventional TAP approach. Assuming that the free energy is self-averaging in this case, we show the equivalence of our TAP approach to the results of a replica calculation.

In Appendix B, an alternative derivation of the TAP free energy is given. It is based on the observation that the functional form (as a function of m_i and M_i) of the Onsager term V_i in the TAP equations does not depend on the specific single variable densities $\rho(\mathbf{S})$. Hence, we may compute the corresponding universal form of $\Delta\Phi$ by calculating Φ for an exactly solvable model, i.e., for a Gaussian ρ and subtract the naive mean-field part. This is the strategy used by Parisi and Potters [30] to derive TAP equations for a spin glass model with an orthogonal random matrix $\mathbf{J}$.

E. Summary of adaptive TAP equations

To summarize our results so far, we write the adaptive TAP equations for $\langle S_i\rangle$ and V_i, $i = 1, \ldots, N$ as

$$\langle S_i\rangle = \frac{\partial}{\partial\theta_i}\ln Z_0^{(i)}, \tag{33}$$

where the single variable partition function is

$$Z_0^{(i)} = \int dS\, \rho_i(S)\exp\left[S\left(\sum_j J_{ij}\langle S_j\rangle - V_i\langle S_i\rangle + \theta_i \right) + \frac{V_i}{2}S^2 \right]. \tag{34}$$

The second set of TAP equations for V_i is

$$\frac{\partial\langle S_i\rangle}{\partial\theta_i} = \frac{\partial^2}{\partial\theta_i^2}\ln Z_0^{(i)} = [(\boldsymbol{\Lambda} - \mathbf{J})^{-1}]_{ii}, \tag{35}$$

where

$$\boldsymbol{\Lambda} = \mathrm{diag}(\Lambda_1,\ldots,\Lambda_N), \quad \Lambda_i \equiv V_i + \left(\frac{\partial\langle S_i\rangle}{\partial\theta_i}\right)^{-1}. \tag{36}$$

Note that the partial derivatives are now taken with respect to the explicit θ_i dependence, i.e., all remaining arguments in $Z_0^{(i)}$ are *fixed*. Finally, the free energy is given by Eq. (32) with $m_i = \langle S_i\rangle$, $M_i = \langle S_i^2\rangle$, $\gamma_i^0 = \langle h_i\rangle_{\setminus i}$, and $\lambda_i^0 = V_i$.

F. The generalized model

The special structure of the augmented coupling matrix (4) allows for a variety of simplifications in treating the model (2). By introducing hatted variables $\langle\hat{\mathbf{S}}\rangle$, $\hat{\mathbf{V}}$, and $\hat{\boldsymbol{\Lambda}}$ explicitly, the previous results for the means (33) and (34) are $\langle S_i\rangle = (\partial/\partial\theta_i)\ln Z_0^{(i)}$ and $\langle\hat{S}_k\rangle = (\partial/\partial\theta_k)\ln\hat{Z}_0^{(k)}$ with

$$Z_0^{(i)} = \int dS\, \rho_i(S) \exp\left[S\left(\sum_j J_{ij}\langle S_j\rangle + \sum_k \hat{J}_{ki}\langle \hat{S}_k\rangle - V_i\langle S_i\rangle + \theta_i \right) + \frac{V_i}{2} S^2 \right], \tag{37}$$

$$\hat{Z}_0^{(k)} = \int d\hat{S}\, \hat{\rho}_k(\hat{S}) \exp\left[\hat{S}\left(\sum_i \hat{J}_{ki}\langle S_i\rangle - \hat{V}_k\langle \hat{S}_k\rangle + \hat{\theta}_k \right) + \frac{\hat{V}_k}{2} \hat{S}^2 \right]. \tag{38}$$

The augmented susceptibility matrix is given by

$$\boldsymbol{\chi}_{\mathrm{aug}} = (\boldsymbol{\Lambda}_{\mathrm{aug}} - \mathbf{J}_{\mathrm{aug}})^{-1} = \begin{pmatrix} \boldsymbol{\Lambda} - \mathbf{J} & -\hat{\mathbf{J}}^T \\ -\hat{\mathbf{J}} & \hat{\boldsymbol{\Lambda}} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\chi} & \tilde{\boldsymbol{\chi}}^T \\ \tilde{\boldsymbol{\chi}} & \hat{\boldsymbol{\chi}} \end{pmatrix},$$

where $\tilde{\chi}_{ki} \equiv \partial(\hat{S}_k)/\partial\theta_i$. It can be shown that although the vector $\hat{\mathbf{S}}$ is purely imaginary, its expectation $\langle\hat{\mathbf{S}}\rangle$ and the susceptibility matrix $\boldsymbol{\chi}_{\mathrm{aug}}$ come out real.

The TAP equations for the V_i's can be simplified using identities for the inverse and the determinant of partitioned matrices [31],

$$\boldsymbol{\chi} = (\boldsymbol{\Lambda} - \mathbf{J} - \hat{\mathbf{J}}^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{J}})^{-1} \tag{39}$$

and $\det(\boldsymbol{\Lambda}_{\mathrm{aug}} - \mathbf{J}_{\mathrm{aug}}) = \det\hat{\boldsymbol{\Lambda}}\, \det(\boldsymbol{\Lambda} - \mathbf{J} - \hat{\mathbf{J}}^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{J}})$, which leads to

$$\hat{\chi}_{kk} = \frac{\partial}{\partial \hat{\Lambda}_k} \ln\det(\boldsymbol{\Lambda}_{\mathrm{aug}} - \mathbf{J}_{\mathrm{aug}}) = \frac{1}{\hat{\Lambda}_k} + \frac{1}{\hat{\Lambda}_k^2} \sum_{ij} \hat{J}_{ki}\hat{J}_{kj}\chi_{ij}, \tag{40}$$

showing that the hatted covariances can be explicitly computed from the nonhatted ones. The variances V_i and $\hat{V}_k$ are obtained from a straightforward generalization of Eq. (35): $\partial\langle S_i\rangle/\partial\theta_i = [(\boldsymbol{\Lambda} - \mathbf{J} - \hat{\mathbf{J}}^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{J}})^{-1}]_{ii}$ and $\partial\langle \hat{S}_k\rangle/\partial\hat{\theta}_k = \hat{\chi}_{kk}$. Finally, we can use the identity for partitioned determinants to show that in the Onsager term (33) a few terms cancel and we can write

$$\Delta\Phi = \tfrac{1}{2}\ln\det(\boldsymbol{\Lambda} - \mathbf{J} - \hat{\mathbf{J}}^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{J}}) - \tfrac{1}{2}\sum_i V_i\chi_{ii} + \tfrac{1}{2}\sum_i \ln\chi_{ii} - \tfrac{1}{2}\sum_k \hat{V}_k\hat{\chi}_{kk} + \tfrac{1}{2}\sum_k \ln(1 + \hat{V}_k\hat{\chi}_{kk}). \tag{41}$$

Finally, we note that for the consistency of the TAP equations, only the positive definiteness of the submatrix $\boldsymbol{\chi}$ for the original real random variables $\mathbf{S}$ (and *not* $\boldsymbol{\chi}_{\mathrm{aug}}$) is required.

III. THERMODYNAMIC LIMIT AND SELF-AVERAGING THEORY

For specific choices of the distribution of disorder, where different sites i appear in a symmetric way, we can expect that quantities like V_i and the free energy become self-averaging in the thermodynamic limit $N\to\infty$, i.e., $V_i = V$ independent of the specific realization of the disorder. Such quantities can be computed by suitable quenched averages using the statistics of the disorder variables J_{ij}.

As usual, quenched averages over the distribution of the coupling matrix $\mathbf{J}$ are performed within the replica framework using $[\ln Z]_J = (d/dn)\ln[Z^n]_{\mathbf{J}}|_{n=0}$, where for integer n we have

$$[Z^n]_{\mathbf{J}} = \int d\mathbf{S}^n \prod_{ia} \rho(S_{ia}) \left[\exp\left(\tfrac{1}{2}\sum_a \sum_{ij} S_{ia} J_{ij} S_{ja} \right) \right]_{\mathbf{J}}. \tag{42}$$

$a = 1, \ldots, n$ are replica indices and the brackets $[\cdots]_{\mathbf{J}}$ denote the average over the J_{ij}'s.

If all matrix elements J_{ij} for $i<j$ are assumed to be iid Gaussian random variables (as for the SK model) the average over $\mathbf{J}$ is easily carried out. The simplest way to generalize this *Gaussian orthogonal ensemble* and allow for correlations between J_{ij}'s is to keep the orthogonality of the ensemble but make it non-Gaussian. Such distributions, which also lead to a well-defined thermodynamic limit (introduced by Ref. [32] and subsequently used by Ref. [30]), are defined by generating functions of the type

$$[e^{1/2\mathrm{Tr}\mathbf{A}\mathbf{J}}]_{\mathbf{J}} = e^{N\ \mathrm{Tr}\ G(\mathbf{A}/N)}, \tag{43}$$

with the function G fully specifying the ensemble. In Appendix D, it is shown how G is related to the spectrum of the matrix $\mathbf{J}$.[2] The Gaussian ensemble is recovered by setting $G(x) \propto x^2$. Note that scaling with $1/N$ inside of G keeps the trace of order one for $N\to\infty$ when the elements of the matrix $\mathbf{A}$ are of order one.

Distributions with generating functions (43) have the nice feature that the average (42) depends only on a single set of quadratic order parameters given by $q_{ab} \equiv (1/N)\Sigma_i S_{ia} S_{ib}$.[3] This can be seen by applying Eq. (43) to the matrix $A_{ij} = \Sigma_{a=1}^n S_{ia} S_{ja}$ appearing in Eq. (42). We note that $\mathbf{A}$ has $N - n$ eigenvalues equal to zero, but in the space spanned by n vectors $\mathbf{S}_a$ we get

$$\frac{1}{N}\sum_j A_{ij} S_{ja} = \sum_a q_{ab} S_{ib}. \tag{44}$$

Hence, in this n-dimensional subspace, the matrix $\mathbf{A}/N$ acts as the matrix $\mathbf{q} = \{q_{ab}\}$.

In this way we can derive general results for the self-averaging case and connect these with the previous results for the adaptive TAP theory. In Secs. III A–III C we compute the Onsager term, the replica free energy, and the average of the TAP free energy and show that both coincide.

[2] In Sec. III A, we discuss the extension of this assumption to the model (2).

[3] Our ensembles do not apply to diluted models, for which usually an infinite sequence of order parameters of any order has to be considered.

Finally, in Sec. III D, we show how the condition of a positive definite susceptibility matrix $\boldsymbol{\chi}$ translates to the AT stability condition.

A. The Onsager term

In this section, we compute $V=[V_i]_{\mathbf{J}}$ and $\Delta\Phi^{\mathrm{self}}=[\Delta\Phi]_{\mathbf{J}}$ from Eq. (32) for the model (1). We will briefly sketch how to generalize these results to the model (2) at the end of the section. To compute V, we write Eq. (35) as

$$\chi_{ii}=[(\mathbf{\Lambda}-\mathbf{J})^{-1}]_{ii}=\frac{\partial}{\partial\Lambda_i}\ln\det(\mathbf{\Lambda}-\mathbf{J}) \tag{45}$$

and replace the right-hand side by its average over the distribution of the matrix $\mathbf{J}$ for $N\to\infty$. Since there is no spin glass ordering for a Gaussian model, it is sufficient to perform an annealed average using the identity

$$\det{}^{-1/2}(\mathbf{\Lambda}-\mathbf{J})=\int\frac{d\mathbf{z}}{(2\pi)^{N/2}}\exp[-\tfrac{1}{2}\mathbf{z}^T(\mathbf{\Lambda}-\mathbf{J})\mathbf{z}]. \tag{46}$$

Applying Eq. (43) to the matrix $\mathbf{A}=\mathbf{z}\mathbf{z}^T$, which has a single eigenvalue equal to $\mathbf{z}^T\mathbf{z}$ (with eigenvector $\mathbf{z}$) and an $(N-1)$-fold degenerate eigenvalue 0, and using the fact that $G(0)=0$ we arrive at

$$\begin{aligned}[\det{}^{-1/2}(\mathbf{\Lambda}-\mathbf{J})]_{\mathbf{J}}&=\int\frac{d\mathbf{z}}{(2\pi)^{N/2}}\exp\left[-\tfrac{1}{2}\sum_i\Lambda_i z_i^2+NG\left(\frac{1}{N}\sum_i z_i^2\right)\right]\\&=\int\frac{dr\,d\hat{r}}{4\pi i/N}\frac{d\mathbf{z}}{(2\pi)^{N/2}}\exp\left[-\tfrac{1}{2}\sum_i\Lambda_i z_i^2+\tfrac{1}{2}\hat{r}\left(\sum_i z_i^2-Nr\right)+NG(r)\right],\end{aligned}$$

where in the last line the order parameter $r=(1/N)\Sigma_i z_i^2$ and its conjugate $\hat{r}$ have been introduced. For $N\to\infty$, $\hat{r}$ is found from the saddle point of

$$\ln\det(\mathbf{\Lambda}-\mathbf{J})=\sum_i\ln(\Lambda_i-\hat{r})+N\hat{r}r-2NG(r), \tag{47}$$

yielding

$$r=\frac{1}{N}\sum_i\frac{1}{\Lambda_i-\hat{r}}. \tag{48}$$

Hence, the averaged TAP equation (35) reads

$$[\chi_{ii}]_J=\frac{1}{\Lambda_i-\hat{r}}, \tag{49}$$

giving

$$r=\overline{\chi}\equiv\frac{1}{N}\sum_i[\chi_{ii}]_J \tag{50}$$

and $V_i=V=\hat{r}$ (when taken together with the definition $\Lambda_i=V_i+1/\chi_{ii}$). Finally, variation with respect to r yields $\hat{r}=2G'(r)$. Summarizing, we find that in the thermodynamic limit,

$$V=2G'(\overline{\chi}). \tag{51}$$

As sketched in Appendix C, the same result is obtained when the general Gaussian model used in the derivation of the Gibbs free energy is replaced by a *spherical* model, where only a single Lagrange parameter λ (or Λ) is coupled to $\Sigma_i S_i^2$.

Inserting the saddle-point values into Eq. (47), the expression for the Onsager term (33) simplifies remarkably,

$$\Delta\Phi^{\mathrm{self}}=NG(\overline{\chi}). \tag{52}$$

Turning to the generalized model, it can be seen from Eqs. (39) and (41) that $\mathbf{J}$ in the original model is replaced by $\mathbf{J}+\hat{\mathbf{J}}^T\hat{\mathbf{\Lambda}}^{-1}\hat{\mathbf{J}}$ in the generalized model. For this case, we extend the definition of orthogonal ensembles (43) to

$$[\exp\{\tfrac{1}{2}\mathrm{Tr}\mathbf{A}(\mathbf{J}+\hat{\mathbf{J}}^T\hat{\mathbf{\Lambda}}^{-1}\hat{\mathbf{J}})\}]_{\mathbf{J},\hat{\mathbf{J}}}=e^{N\,\mathrm{Tr}G_{\hat{\mathbf{\Lambda}}}(\mathbf{A}/N)}. \tag{53}$$

With this definition, the result (51) for V remains valid, i.e., $V=2G'_{\hat{\mathbf{\Lambda}}}(\overline{\chi})$. Self-averaging results for $\hat{V}_k$ obtained from Eq. (39) and the Onsager term (41) are derived for neural network models in Sec. IV.

B. Replica free energy

To get the average free energy, we use Eq. (42) and compute[4]

$$\begin{aligned}[Z^n]_{\mathbf{J}}&=\int dS^n\prod_{ia}\rho(S_{ia})\left[\exp\left(\tfrac{1}{2}\sum_a\sum_{ij}S_{ia}J_{ij}S_{ja}\right)\right]_{\mathbf{J}}\\&=\int d\mathbf{S}^n\prod_{ia}\rho(S_{ia})e^{N\,\mathrm{Tr}G(\mathbf{q})}.\end{aligned} \tag{54}$$

In replica symmetry, $\mathbf{q}$ has only two types of eigenvalues: a nondegenerate one given by $(n-1)q+q_0$ and an $(n-1)$-fold degenerate eigenvalue equal to q_0-q. Hence

$$\mathrm{Tr}\,G(\mathbf{q})=(n-1)G(q_0-q)+G(nq+[q_0-q]). \tag{55}$$

After introducing and eliminating conjugate parameters $\hat{q}$ and $\hat{q}_0$ with a saddle-point method, we find the replica symmetric free energy

[4]This analysis can be easily generalized to include cases with more than one order parameter, e.g., for a neural network learning from a teacher.

$$-\frac{1}{N}[\ln Z]_{\mathbf{J}}=-\frac{1}{N}\frac{d}{dn}\ln[Z^n]_{\mathbf{J}}\Big|_{n=0}$$

$$=-\int Dz\ln\int dS\,\rho(S)\exp[\sqrt{2qG''(q)}zS+G'(q_0-q)S^2]-G(q_0-q)+(q_0-q)\times[qG''(q_0-q)+G'(q_0-q)], \quad (56)$$

where $Dz\equiv dze^{-z^2/2}/\sqrt{2\pi}$ and q_0 and q are obtained from the saddle point.

C. Averaging the TAP free energy

We will next prove the consistency of our TAP approach with the results of the replica theory in the thermodynamic limit. We want to show that the averaged free energy (56) calculated with the help of replicas coincides with the disorder averaged Gibbs free energy from the TAP approximation. From this result we can conclude that self-averaging quantities that can be derived from the free energy by derivatives with respect to external fields will be exact in the TAP approach.

In order to prove this, we define an auxiliary partition function, which reproduces the TAP Gibbs free energy $\Phi_1(\mathbf{m},\mathbf{M})$, Eq. (32), evaluated at equilibrium, i.e., for the values of $\mathbf{m}$ and $\mathbf{M}$ provided by the solutions of the TAP equations. We will only give a brief description. The calculation uses the representations (28) and (29) of the free energy. With Eq. (32) and the thermodynamic limit simplifications (52), $\lambda_i\to\lambda$, and defining $q_0=(1/N)\Sigma_i M_i$, we rewrite Eq. (29) as

$$\Psi=-\ln Z(\lambda\mathbf{I},\boldsymbol{\gamma})+\sum_i m_i\gamma_i+\frac{\lambda}{2}q_0N-\frac{1}{2}\sum_{ij}m_iJ_{ij}m_j+NG\left(q_0-\frac{1}{N}\sum_i m_i^2\right).$$

Following Eq. (28), Ψ will coincide with the equilibrium value of the Gibbs free energy when evaluated at the values for $\boldsymbol{\gamma}$, $\mathbf{m}$, λ, and q_0, which make Ψ *stationary*. The stationary values are equal to those obtained from the TAP equations when we identify $m_j\equiv\langle S_j\rangle$ and $q_0=(1/N)\Sigma_i\langle S_i^2\rangle$. The auxiliary partition function

$$Y=\int d\boldsymbol{\gamma}\,d\mathbf{m}e^{-\beta\Psi} \quad (57)$$

is in the limit $\beta\to\infty$ dominated by the values of m_i and γ_i for which Ψ is stationary, provided the paths of integration are chosen such that the integral exists.[5] Assuming also stationarity with respect to λ and q_0, we recover the TAP free energy at equilibrium from

$$\Phi=-\operatorname*{extr}_{\lambda,q_0}\lim_{\beta\to\infty}\frac{1}{\beta}\ln Y. \quad (58)$$

Variation with respect to q_0 yields $\lambda=2G'(q_0-(1/N)\Sigma_i m_i^2)$. The calculation of $[\Phi]_{\mathbf{J}}$ proceeds by a straightforward replica calculation using the average

$$\frac{1}{N}\ln\left[\exp\left(\frac{1}{2}\sum_a\sum_{ij}\beta m_{ia}J_{ij}m_{ja}\right)\right]_{\mathbf{J}}=(n-1)G(\beta(q-\overline{q}))+G(\beta(n\overline{q}+q-\overline{q})). \quad (59)$$

For $\beta\to\infty$, the integrations over γ_i and m_i are decoupled and performed by the saddle-point method yielding $m_i=0$ and $\gamma_i=\sqrt{qG''(q-q_0)}z_i$, where z_i is a standard normal random variable showing the equivalence to Eq. (56). In comparing both replica calculations, it is useful to note that by a linear response argument we can identify

$$\lim_{\beta\to\infty}\beta(q-\overline{q})=\overline{\chi}=q-q_0. \quad (60)$$

Putting everything together we find that $[\Phi]_{\mathbf{J}}=-[\ln Z]_{\mathbf{J}}$.

D. Stability and AT condition

We will show next (by generalizing the arguments of [33]) how the positive definiteness of the susceptibility matrix (i.e., the matrix of covariances) $\boldsymbol{\chi}$, Eq. (24) [or Eq. (39) for the model (2)], translates into the de Almeida–Thouless stability condition well known from the replica theory.

From Eq. (24), positive definiteness of $\boldsymbol{\chi}$ is equivalent to the condition that $\mathbf{H}=\boldsymbol{\Lambda}-\mathbf{J}$ has only positive eigenvalues. Hence, in the thermodynamic limit, the eigenvalue density $\rho(\gamma)\equiv\lim_{N\to\infty}(1/N)\Sigma_\mu\delta(\mu-\gamma)$, where μ denotes the eigenvalues of $\mathbf{H}$, must be exactly zero for small positive γ. Using a standard representation of δ functions, we have

$$\rho(\gamma)=\frac{1}{N\pi}\lim_{\delta\to0^+}\mathrm{Im}\sum_\mu\frac{1}{\mu-\gamma-i\delta}=\frac{1}{N\pi}\lim_{\delta\to0^+}\mathrm{Im}\,\mathrm{Tr}[\mathbf{H}-(\gamma+i\delta)\mathbf{I}]^{-1}=-\frac{1}{N\pi}\lim_{\delta\to0^+}\mathrm{Im}\frac{\partial}{\partial\gamma}\ln\det[\boldsymbol{\Lambda}-\mathbf{J}-(\gamma+i\delta)\mathbf{I}].$$

Since we have already calculated $\ln\det(\boldsymbol{\Lambda}-\mathbf{J})$ in Sec. III A, we can immediately write down the result as

$$\rho(\gamma)=\lim_{\delta\to0^+}\mathrm{Im}\frac{1}{\pi N}\sum_i\frac{1}{\Lambda_i-\hat{r}(\gamma)-(\gamma+i\delta)}. \quad (61)$$

In general, it is hard to obtain a closed form solution to the saddle-point equation for $\hat{r}(\gamma)$. However for γ close to zero—the interesting region with regard to the stability—$\hat{r}(\gamma)$ is close to $\hat{r}(0)=V$ and we can easily get a solution for

[5]Note that the stationary value of Ψ is not given by a minimum but by a saddle point.

$\hat{r}$ and r by expanding to second order in $\delta\hat{r} \equiv \hat{r}(\gamma) - \hat{r}(0)$ and $\delta r = r(\gamma) - r(0), r(0) = \overline{\chi}$. We find that as long as the condition

$$1 - 2G''(\overline{\chi}) \frac{1}{N} \sum_i [\chi_{ii}]^2_{\mathbf{J}} > 0 \tag{62}$$

is satisfied $\hat{r}(\gamma)$ is *real* and the density $\rho(\gamma)$, Eq. (61), vanishes for small $\gamma > 0$.

On the other hand, if the left-hand side of Eq. (62)—the stability condition—is zero, then $\delta\hat{r}$ has an imaginary part

$$\delta\hat{r} = \left\{ \frac{-\gamma}{2G''(\overline{\chi}) \frac{1}{N} \sum_i [\chi_{ii}]^3_{\mathbf{J}} + \frac{G'''(\overline{\chi})}{[2G''(\overline{\chi})]^2}} \right\}^{1/2} \tag{63}$$

and the support of the density of eigenvalues, Eq. (61),

$$\rho(\gamma) \simeq \frac{1}{\pi N} \sum_i [\chi_{ii}]^2_{\mathbf{J}} \, \mathrm{Im}\, \delta\hat{r} = \frac{1}{\pi} \left\{ \frac{\gamma}{[2G''(\overline{\chi})]^3 \frac{1}{N} \sum_i [\chi_{ii}]^3_{\mathbf{J}} + G'''(\overline{\chi})} \right\}^{1/2} \tag{64}$$

extends to $\gamma = 0$ and the solutions of TAP equations are only marginally stable. As we will show for some examples in the next section, Eq. (62) coincides with the AT stability condition of replica theory [1].

IV. APPLICATIONS

In this section we will give explicit examples for adaptive TAP equations for a few models that have been previously considered in the literature. These are the SK and Hopfield models and the perceptron. The latter is of the generalized model type (2). Finally, in simulations, we investigate the effect of the choice of the Onsager term for perceptron learning problems.

A. Ising models

For Ising models we have a prior distribution $\rho(S) = \frac{1}{2}\delta(S-1) + \frac{1}{2}\delta(S+1)$ so that

$$Z_0^{(i)} = \cosh\left(\sum_{j,j\neq i} J_{ij} \langle S_j \rangle - V_i \langle S_i \rangle + \theta_i \right),$$

which leads to

$$\langle S_i \rangle = \tanh\left(\sum_{j,j\neq i} J_{ij} \langle S_j \rangle - V_i \langle S_i \rangle + \theta_i \right)$$

and $\chi_{ii} = 1 - \langle S_i \rangle^2$.

1. SK model

For the SK model [3] the statistics of the couplings is given by $\overline{J_{ij}} = 0$, $\overline{J_{ij}^2} = \beta/N$ with $J_{ij} = J_{ji}$. Then the G function (43) becomes $G(r) = (\beta r)^2/4$ and according to Eq. (51),

$$V = 2G'(1-q) = \beta^2(1-q), \tag{65}$$

where $q = (1/N)\Sigma_i \langle S_i \rangle^2$ is the *Edwards Anderson* parameter. The stability condition simplifies to $1 - (\beta^2/N)\Sigma_i \chi_{ii}^2 > 0$ and $\rho(\gamma) = (1/\pi)\sqrt{\gamma/(\beta^6/N)\Sigma_i \chi_{ii}^3}$ in agreement with [33].

2. Hopfield model

The coupling matrix of the Hopfield model [17] is $J_{ij} = (\beta/N)\Sigma_k^m x_{ki} x_{kj}$, where we assume that the x_{ki} are iid random variables of zero mean and unit variance. The G function[6] (43) is then found to be

$$G(r) = -\frac{m}{2N}[\ln(1-\beta r) + \beta r], \tag{66}$$

leading to

$$V = 2G'(1-q) = \frac{m}{N} \frac{\beta^2(1-q)}{1-\beta(1-q)}, \tag{67}$$

in agreement with [1].

B. Perceptron

Perceptrons are single layer neural networks that are parametrized by a vector of weights $\mathbf{S}$. We consider both the learning of regression and binary classification problems from a training set that is given by $\{(\mathbf{x}_k, y_k), k = 1, \ldots, m\}$. $\mathbf{x} \in R^N$ denotes a vector of inputs and $y \in \mathcal{R}$ is a real valued output for regression and a binary label $y = \pm 1$ for classification. In the first case the output of the perceptron is given by $\mathbf{S}\cdot\mathbf{x}$ and in the latter case by $\mathrm{sgn}(\mathbf{S}\cdot\mathbf{x})$. Although this simple linear model is of limited power compared to multilayer neural networks, it can be easily generalized to the so-called *Gaussian process models*. These are able to make nonlinear predictions and achieve state-of-the-art performance on a variety of standard benchmark data sets. An application of the adaptive TAP approach to the Gaussian process models was given in [8].

Perceptrons can be understood as probabilistic models by defining a probability (likelihood) $P(y|\mathbf{S}\cdot\mathbf{x})$ for the observations y given inputs $\mathbf{x}$ and weights $\mathbf{S}$. For classification we consider the so-called *probit* model, which can be derived by assuming that labels are generated as $y = \mathrm{sgn}(\mathbf{S}\cdot\mathbf{x} + u)$, where u is a Gaussian noise of variance σ^2. Hence, $P(y|\mathbf{S}\cdot\mathbf{x}) = \phi(y[\mathbf{S}\cdot\mathbf{x}/\sigma])$, where $\phi(z) \equiv \int_{-\infty}^{z} Dt$. In the noise-free limit ϕ reduces to the unit step function. For regression with additive Gaussian noise the likelihood is $P(y|\mathbf{S}\cdot\mathbf{x}) \propto \exp[-(y - \mathbf{S}\cdot\mathbf{x})^2/2\sigma^2]$.

The model is clearly of the form given by Eq. (2) with $\hat{J}_{ki} = x_{ki}$ and $\mathbf{J} = \mathbf{0}$. We identify the likelihood $P(y|\hat{h})$ with $F(\hat{h})$ in Eq. (3) where $\hat{h} = \mathbf{S}\cdot\mathbf{x}$. The explicit appearance of the hatted variables in the algorithm will be especially useful when we want to discuss the important effects of removing a

[6] This is easily shown for Gaussian x_{ki}. For binary x_{ki}, the average must be restricted to the condensed patterns and the relation (43) will hold only for $N \to \infty$.

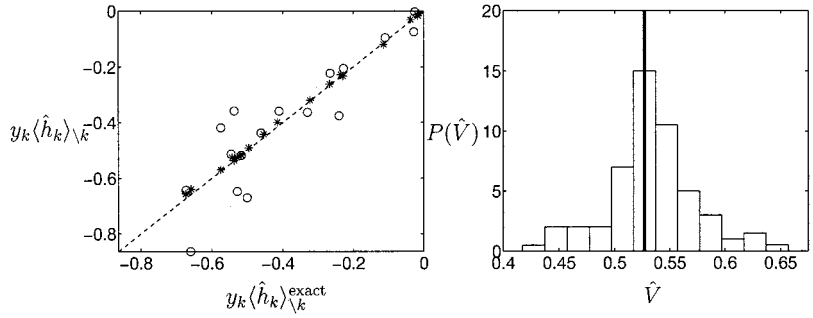

FIG. 1. Test of self-consistency of TAP—$y_k(\hat{h}_k)_{\setminus k}$ versus $y_k(\hat{h}_k)_{\setminus k}^{\text{exact}}$. The stars/circles are for adaptive/conventional TAP. The right plot shows the distribution of the cavity variances $\hat{V}_k$. The line in the middle is the value found from the self-averaging theory. The plot is for the noise-free perceptron with $m=N=100$.

data point (rather than a weight S_j) from the set of training examples.

Under the assumptions of a Gaussian cavity field, the Bayes predictors for regression and classification become $\langle\hat{h}\rangle=\langle\mathbf{S}\rangle\cdot\mathbf{x}$ and $\text{sgn}\langle\hat{h}\rangle$ [12]. For the weight variables S_j we consider both Ising weights $\rho(S)=\frac{1}{2}\delta(S-1)+\frac{1}{2}\delta(S+1)$ and weights with a Gaussian prior distribution $\rho(S)=e^{-S^2/2}/\sqrt{2\pi}$. In the first case, we recover the Ising result

$$\langle S_i\rangle=\tanh\left(\sum_k x_{ki}\langle\hat{S}_k\rangle-V_i\langle S_i\rangle+\theta_i\right), \tag{68}$$

and in the Gaussian case we simply get

$$\langle S_i\rangle=\sum_k x_{ki}\langle\hat{S}_k\rangle+\frac{\theta_i}{1-V_i}.$$

The TAP equations for the hatted variables $\langle\hat{S}_k\rangle=\partial\ln\hat{Z}_0^{(k)}/\partial\hat{\theta}_k$ are obtained from Eq. (38),

$$\hat{Z}_0^{(k)}=\int DzP(y_k|\langle\hat{h}_k\rangle_{\setminus k}+\hat{\theta}_k+\sqrt{\hat{V}_k}z),$$

with $\langle\hat{h}_k\rangle_{\setminus k}=\Sigma_i x_{ki}\langle S_i\rangle-\hat{V}_k\langle\hat{S}_k\rangle$. Explicit expressions are for classification,

$$\hat{Z}_0^{(k)}=\phi\left(y_k\frac{\langle\hat{h}_k\rangle_{\setminus k}+\hat{\theta}_k}{\sqrt{\sigma^2+\hat{V}_k}}\right),$$

and for regression,

$$\hat{Z}_0^{(k)}=\frac{1}{\sqrt{2\pi(\sigma^2+\hat{V}_k)}}\exp\left[-\frac{(y_k-\langle\hat{h}_k\rangle_{\setminus k}-\hat{\theta}_k)^2}{2(\sigma^2+\hat{V}_k)}\right].$$

To connect with results known in the literature, we derive the self-averaging properties for the case, where the x_{ki} are iid random variables with zero means and variance $1/N$. The G function (53) becomes

$$G_{\hat{\Lambda}}(r)=-\frac{1}{2N}\sum_k\ln(1-r/\hat{\Lambda}_k).$$

The self-averaging value for the variances of the original variables $V_i=V$ is given by Eq. (51) and the variance for the hatted variables is given by Eq. (40). Taken together they lead to the symmetric result in the two sets of variables

$$\hat{V}=\overline{\chi}=\frac{1}{N}\sum_i[\chi_{ii}]_{\mathbf{J},\hat{\mathbf{J}}}, \tag{69}$$

$$V=\overline{\hat{\chi}}\equiv\frac{1}{N}\sum_k[\hat{\chi}_{kk}]_{\mathbf{J},\hat{\mathbf{J}}}. \tag{70}$$

The Onsager term (41), stability condition (62), and the eigenvalue spectrum (64) become

$$\Delta\Phi^{\text{self}}=-\frac{N}{2}V\hat{V}, \tag{71}$$

$$1-\frac{1}{N^2}\sum_k\hat{\chi}_{kk}^2\sum_i\chi_{ii}^2>0, \tag{72}$$

$$\rho(\gamma)\simeq\frac{1}{\pi}\left\{\frac{\gamma}{\left[\frac{1}{N}\sum_k\hat{\chi}_{kk}^2\right]^3\frac{1}{N}\sum_i\chi_{ii}^3+\frac{1}{N}\sum_k\hat{\chi}_{kk}^3}\right\}^{1/2}. \tag{73}$$

Specializing to a Gaussian weight prior for which $\chi_{ii}=1/(1-V_i)$, we find in accordance with previous results [10–12] that $\hat{V}=1/(1-V)$ and the stability condition reduces to $1-\hat{V}^2(1/N)\Sigma_k\hat{\chi}_{kk}^2>0$.

Finally, we test the adaptive and self-averaging TAP equations in two learning scenarios. We first test the internal consistency of the theory by comparing the cavity field calculated from the solution of the TAP equations $\langle\hat{h}_k\rangle_{\setminus k}=\Sigma_i x_{ki}\langle S_i\rangle-\hat{V}_k\hat{S}_k$ with the ''exact'' cavity field $\langle\hat{h}_k\rangle_{\setminus k}^{\text{exact}}$ computed by actually *removing* example k from the training set and solving the TAP equations for the remaining $m-1$ examples and repeating this procedure for $k=1,\ldots,m$. A

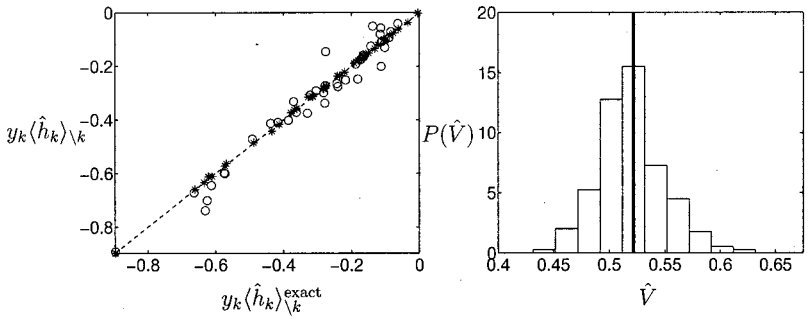

FIG. 2. The same as above with $m=N=200$.

precise estimate of the cavity field is of practical relevance in machine learning since it can be used to define ''leave-one-out'' estimators of the generalization error [12,8,22]. One such 'leave-one-out' estimate for classification is the fraction of negative terms $y_k\langle\hat{h}_k\rangle_{\setminus k}$ over the training set: $\epsilon_{\text{loo}}=(1/m)\Sigma_k\Theta(-y_k\langle\hat{h}_k\rangle_{\setminus k})$ since $\text{sgn}\langle\hat{h}_k\rangle_{\setminus k}$ is the leave-one-out prediction of y_k.

Figures 1 and 2 show the result of learning in the simple perceptron with Gaussian weight prior, the likelihood for noise-free classification, and the iid distribution of inputs for which the self-averaging theory is expected to become exact in the thermodynamic limit. The output labels are generated by a neural network teacher $y=\text{sgn}(\mathbf{T}\cdot\mathbf{x})$. While for positive values of $y_k\langle\hat{h}_k\rangle_{\setminus k}$ (not shown in the figures), i.e., the examples for which the leave-one-out prediction is correct, the agreement between $\langle\hat{h}_k\rangle_{\setminus k}$ and $\langle\hat{h}_k\rangle_{\setminus k}^{\text{exact}}$ tends to be better, the negative values are more crucial for real applications because they give the desired leave-one-out error count. The results clearly show that the internal consistency of the adaptive TAP is better than that of the self-averaging theory. The results indicate that finite size effects are quite important even for reasonably large systems and that the adaptive theory is better at taking these into account.

Performing the same analysis for real data gives even more striking results. Here we consider the data set ''Sonar—Mines versus Rocks'' [34] of size $m=104$ with binary class labels $y_k=\pm1$ and a $N=60$-dimensional input space. We use the Gaussian prior for the weights and $\sigma^2=0.5$ in the likelihood. In Fig. 3 we again plot $y_k\langle\hat{h}_k\rangle_{\setminus k}$ versus $y_k\langle\hat{h}_k\rangle_{\setminus k}^{\text{exact}}$. For the adaptive theory, we find a perfect agreement between the two computations of the leave-one-out estimate: $\epsilon_{\text{loo}}=\epsilon_{\text{loo}}^{\text{exact}}=\frac{33}{104}$. For comparison, the self-averaging TAP approach gives $\epsilon_{\text{loo}}=\frac{41}{104}$ and $\epsilon_{\text{loo}}^{\text{exact}}=\frac{33}{104}$. The consistency of the leave-one-out error based on the adaptive TAP approach is also apparent in the generalization of the perceptron to the Gaussian process models (see [8]).

In the second set of simulations we have tested the influence of using a wrong cavity variance V in the mean-field equations (and wrong Onsager term $\Delta\Phi$ in the free energy). Since the free energy is the negative log likelihood of the observed data, i.e., *at equilibrium*, $\Phi=-\ln P(\mathbf{y})$, it can be used for deciding which model gives the best fit to data. It is therefore also of practical interest to get a reliable estimate of Φ.

In these simulations we consider a nontrivial regression model with binary weights. See Ref. [35] for a discussion of this model in the context of demodulation in communications systems. As it can be seen directly from the likelihood, the regression problem can alternatively be regarded as an N-dimensional model of the type (1), i.e., $\Pi_k P(y_k|\mathbf{S}\cdot\mathbf{x}_k)\propto\exp(\Sigma_{i>j}S_iJ_{ij}S_j+\Sigma_iS_i\theta_i)$. The couplings and external fields are given by $J_{ij}=-\Sigma_k x_{ki}x_{kj}/\sigma^2$ and $\theta_i=\Sigma_k x_{ki}y_k/\sigma^2$. We can now directly compare the use of the correct self-averaging V, Eq. (70), for this model with that provided from the adaptive TAP approach and that of other Ising models with different random matrix ensembles, namely, the SK and Hopfield model equations (65) and (67).

In Fig. 4, we compare the TAP mean-field free energy Φ found in simulations using the different expressions for the Onsager term with the prediction of replica theory [36]. In the simulations $N=60$, $\sigma^2=0.2$, and the training set is generated by a noise-free binary teacher: $y=\mathbf{T}\cdot\mathbf{x}$ with $T_i=\pm1$. The simulations are averaged over 100 runs and the error bars are of the size of the symbols. We set $\beta=1$ for the SK model and $\beta=0.99$ for the Hopfield model.[7] The figure shows that both adaptive and self-averaging TAP results with the Onsager term (71) are in excellent agreement with replica theory. Using the SK and Hopfield-Onsager terms tends to produce saturated solutions, i.e., $\langle S_i\rangle\approx\pm1$ even for training set sizes where this is not expected theoretically, and leads to a completely wrong estimate of the free energy.

V. SUMMARY AND OUTLOOK

We have presented a generalization of the TAP approach for disordered systems, which is able to cope with the lack of knowledge of the disorder distribution. Such a generalization is necessary for the recent applications of mean-field methods to probabilistic data models.

We have demonstrated the significance of our approach in two ways: We have shown that it reproduces the correct thermodynamic limit results for a class of disorder distributions compatible with fully connected models in replica symmetry. Second, the application of our approach to toy models as well as real data models has shown the importance of

[7]The latter was chosen in order to avoid numerical problems when $q\approx1$.

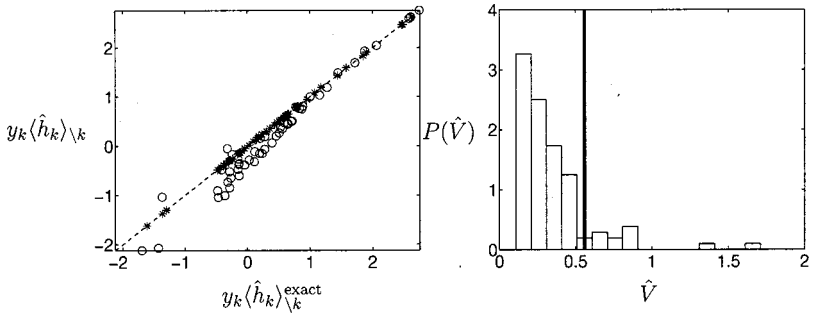

FIG. 3. The same as above for the Sonar data set. For clarity we have left out half of the data points in the left plot.

using the correct TAP approximation when good approximations for *leave-one-out estimators* of errors and *free energies* are required. Such quantities serve as practical yardsticks for comparing different data models and assessing the validity of model predictions.

While the present framework may be sufficient for a variety of practical applications, it could nevertheless break down when the probabilistic model is multimodal having many modes of almost equal weights (free energies). For systems in such ''glassy'' phases our TAP solutions are expected to violate the stability condition and an extension of our framework to a theory, which includes Parisi's hierarchical organization of states, would be highly important. We expect that this is possible by generalizing the ideas presented in Chap. V of [1]. However, one may speculate that in such cases, solving the TAP equations may be highly nontrivial.

We conclude with two other problem areas that have high priority for our future research. These are the limitations of our method to models with extensive connectivities and the algorithmic aspects of our approach, i.e., the development of efficient algorithms for solving the TAP equations. Recent studies on other advanced mean-field techniques indicate that both problems have interesting relationships and also give promising directions for their solutions.

The so-called belief propagation algorithm [6], used in the field of artificial intelligence for approximate probabilistic computations on graphical models with sparse connectivities, was recently identified [15,37,38] as an efficient method to solve the Bethe approximation (a cavity type of approximation) of statistical physics. This observation has already led to principled ways of combining the improved accuracy of higher order (Kikuchi) Bethe approximations [38] with the efficiency of the belief propagation method.

Another interesting approach to an approximate propagation of probability distributions when data arrive sequentially is the Bayesian on-line method introduced in [39,40] and further developed in [41–43]. This technique can be formulated for fairly general model classes but was so far limited to a single sweep through the data, thereby making the approximation dependent on the ordering of the data sequence. In a recent study by Minka [44] it was shown that by a proper recycling of the data, a convergence to the solutions of the TAP equations for the case of a Gaussian process classifier [8] was achieved. We expect that by a consequent and principled combination of the cavity idea with algorithms that are similar to the on-line or the belief propagation technique, we will not only get efficient methods for the solution of TAP equations but also be able to significantly extend the range of applications of the TAP approach.

ACKNOWLEDGMENTS

This research is supported by the Swedish Foundation for Strategic Research as well as the Danish Research Councils through the THOR Center for Neuroinformatics and Center for Biological Sequence Analysis.

APPENDIX A: CORRECTNESS OF TAP EQUATIONS FOR THE GAUSSIAN MODEL

We will show that the relation (24),

$$\boldsymbol{\chi}=(\boldsymbol{\Lambda}-\mathbf{J})^{-1}, \tag{A1}$$

is correct for a Gaussian model. Gaussian models are defined by $\rho_i(S_i)\propto e^{-(1/2)A_i x_i^2}$ and we have always $\boldsymbol{\chi}=(\mathbf{A}-\mathbf{J})^{-1}$, where $\mathbf{A}=\mathrm{diag}(A_1,\ \ldots,\ A_N)$. Hence, we only have to show that $A_j=\Lambda_j=V_j+1/\chi_{jj}$. To see this we look at the single variable partition function (18) derived from the Gaussian cavity field assumption, which is exact for the Gaussian model

$$Z_0^{(i)}=\int dS\ \rho_i(S)\exp[S(\langle h_i\rangle_{\setminus i}+\theta_i)+\tfrac{1}{2}V_iS^2]. \tag{A2}$$

This gives in fact $\chi_{ii}=\partial^2\ln Z_0^{(i)}/\partial\theta_i^2=1/(A_i-V_i)$.

APPENDIX B: ADAPTIVE TAP FREE ENERGY II

Parisi and Potters [30] in their analysis of a spin glass model with random orthogonal couplings made the important observation—motivated by a high temperature expansion—that (within the TAP approximation) two models having the same interactions $\Sigma_{i<j}S_iJ_{ij}S_j$ but differing only in their single spin constraints $\rho_i(S)$, should have free energies Φ that differ also only in the ''single variable'' (or entropic) contribution Φ_0, Eq. (31).

Hence, it is possible to compute the TAP approximation for the free energy Φ from the free energy for an exactly solvable model Φ^s, the entropic term for the solvable model Φ_0^s, and the single variable term for our model Φ_0, i.e.,

$$\Phi=\Phi^s-\Phi_0^s+\Phi_0. \tag{B1}$$

Since the TAP equations for a Gaussian model are exact (see Appendix A) we choose $\rho(S)=e^{-S^2/2}/\sqrt{2\pi}$ for the solvable

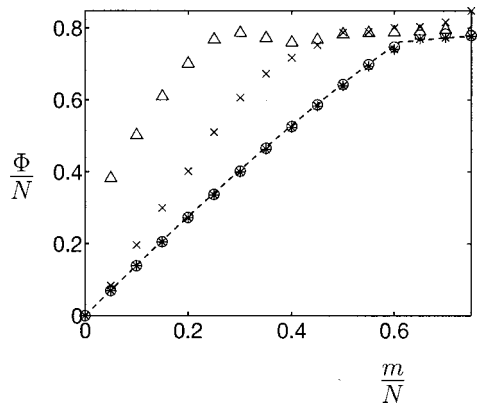

FIG. 4. The free energy Φ as a function of the training set size m. The dashed line is the prediction of replica theory. Stars/circles (almost coinciding) are the results for adaptive TAP/correct self-averaging TAP. Crosses/triangles are the results for self-averaging TAP with the Hopfield/SK Onsager term.

model. We easily get the following exact result for its Gibbs free energy (after eliminating the Lagrange multipliers γ_i and with hindsight redefining $\Lambda_i \equiv 1-\lambda_i$):

$$\Phi^s(\mathbf{m},\mathbf{M}) = \tfrac{1}{2}\ln\det(\mathbf{\Lambda}-\mathbf{J}) + \tfrac{1}{2}\sum_{ij} m_i J_{ij} m_j - \tfrac{1}{2}\sum_i \chi_{ii}\Lambda_i + \sum_i \frac{M_i}{2}, \tag{B2}$$

where we have to insert the value for Λ_i, which solves

$$\chi_{ii} = \langle S_i^2\rangle - \langle S_i\rangle^2 = [(\mathbf{\Lambda}-\mathbf{J})^{-1}]_{ii}. \tag{B3}$$

The single variable term for the Gaussian model Φ_0^s is found by setting $J_{ij}=0$ in Eq. (B2). Eliminating $\mathbf{\Lambda}$ using $\partial_{\mathbf{\Lambda}}\Phi_0^s = 0$,

$$\Phi_0^s(\mathbf{m},\mathbf{M}) = -\tfrac{1}{2}\sum_i \ln\chi_{ii} - \frac{N}{2} + \sum_i \frac{M_i}{2}. \tag{B4}$$

We can now write down the general result for the TAP mean-field Gibbs free energy for a model of the type (1). Collecting the terms in Eq. (B2) and (B4), we arrive at the free energy

$$\Phi = \Phi^s - \Phi_0^s + \Phi_0 = \Phi_0 - \tfrac{1}{2}\sum_{ij} m_i J_{ij} m_j + \Delta\Phi, \tag{B5}$$

$$\Delta\Phi = \tfrac{1}{2}\ln\det(\Lambda - \mathrm{J}) - \tfrac{1}{2}\sum_i \Lambda_i\chi_{ii} + \tfrac{1}{2}\sum_i \ln\chi_{ii} + \frac{N}{2}. \tag{B6}$$

This result should be compared to Eq. (32). Using the saddle-point condition $\partial_{M_i}\Phi = 0$, which implies $\Lambda_i = 1/\chi_{ij} + \lambda_i$, we can rewrite the Onsager term in the form of Eq. (33), where it should be noted that $V_i = \lambda_i$. We have thus rederived the result obtained in Sec. II D.

APPENDIX C: FREE ENERGY FOR THE SPHERICAL MODEL

For the spherical model defined by the constraint $\Sigma_i S_i^2 = N$ we obtain

$$\Phi = \Phi_0 - \tfrac{1}{2}\sum_{ij}\langle S_i\rangle J_{ij}\langle S_j\rangle + \tfrac{1}{2}\ln\det(\Lambda\mathbf{I}-\mathbf{J}) - \frac{N\Lambda\chi}{2} + \frac{N}{2}\ln\chi - \frac{N}{2}, \tag{C1}$$

where $\chi \equiv (1/N)\Sigma_i\chi_{ii}$ and there is also only a single λ^0 in Φ_0. Λ is determined by

$$\chi = \frac{1}{N}\mathrm{Tr}(\Lambda\mathbf{I}-\mathbf{J})^{-1}. \tag{C2}$$

Second, we have $\Lambda = 1/\chi + V$. Repeating the same averaging step as before yields again Eq. (51).

APPENDIX D: EIGENVALUE SPECTRUM OF J

In this appendix we will show how the G function (43) can be expressed in terms of the eigenvalue spectrum of the matrix $\mathbf{J}$. We define

$$r \equiv \int d\mu \frac{p(\mu)}{\Lambda-\mu} = \frac{1}{N}\mathrm{Tr}(\Lambda\mathbf{I}-\mathbf{J})^{-1}, \tag{D1}$$

where $p(\mu)$ is the density of eigenvalues of $\mathbf{J}$. By adding a small imaginary part to Λ we get apart from a factor, directly the density. Using again the Gaussian representation of the determinant yields the equations $r = 1/(\Lambda - \hat{r})$, where $\hat{r}$ is the order-parameter conjugate to $\Sigma_i z_i^2$. It obeys $\hat{r} = 2G'(r)$. Hence

$$\frac{1}{2r} - \frac{\Lambda}{2} + G'(r) = 0. \tag{D2}$$

Solving Eq. (D2) enables us to compute $r(\Lambda)$ when the function $G(r)$ is given. We may also get G as a function of Λ by integrating Eq. (D2).

[1] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, Lecture Notes in Physics Vol. 9 (World Scientific, Singapore, 1987).
[2] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philos. Mag. **35**, 593 (1977).
[3] D. Sherrington and K. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[4] C. De Dominicis, Phys. Rep. **67**, 37 (1980).

[5] *Learning in Graphical Models*, edited by M. Jordan (MIT Press, Cambridge, MA, 1999).
[6] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Kaufmann, San Francisco, 1988).
[7] T.-W. Lee, *Independent Component Analysis* (Kluwer Academic, Boston, 1998).
[8] M. Opper and O. Winther, Neural Comput. **12**, 2655 (2000).
[9] G. E. Hinton and T. J. Sejnowski, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Washington, 1983 (IEEE Press, New York, 1983), p. 448.
[10] M. Mézard, J. Phys. A **22**, 2181 (1989).
[11] K. Y. M. Wong, Europhys. Lett. **30**, 245 (1995).
[12] M. Opper and O. Winther, Phys. Rev. Lett. **76**, 1964 (1996).
[13] H. J. Kappen and F. B. Rodríguez, Neural Comput. **10**, 1137 (1998).
[14] T. Tanaka, Phys. Rev. E **58**, 2302 (1998).
[15] Y. Kabashima and D. Saad, Europhys. Lett. **44**, 668 (1998).
[16] *Advanced Mean Field Methods, Theory and Practice*, edited by M. Opper and D. Saad (MIT Press, Cambridge, MA, 2001).
[17] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[18] T. Plefka, J. Phys. A **15**, 1971 (1982).
[19] C. Bhattacharyya and S. S. Keerthi, J. Phys. A **10**, 1307 (2000).
[20] S. Amari, S. Ikeda, and H. Shimokawa, in *Advanced Mean Field Methods Theory and Practice* (Ref. [16]).
[21] H. J. Kappen and W. Wiegerinck, in *Advanced Mean Field Methods, Theory and Practice* (Ref. [16]).
[22] M. Opper and O. Winther, Phys. Rev. Lett. **86**, 3695 (2001).
[23] M. Mézard and G. Parisi, J. Phys. (France) Lett. **46**, L771 (1985).
[24] D. J. C. MacKay, Maximum Likelihood and Covariant Algorithms for Independent Component Analysis, University of Cambridge, Cavendish Laboratory, Draft 3.7, 1996 (unpublished).
[25] R. Neal, Artif. Intel. **56**, 71 (1992).
[26] L. K. Saul, T. Jaakkola, and M. I. Jordan, J. Artif. Intell. Res. **4**, 61 (1996).
[27] M. Ohlsson, C. Peterson, and B. Söderberg, Neural Comput. **5**, 331 (1993).
[28] Y. Kabashima and D. Saad, in *Advanced Mean Field Methods, Theory and Practice* (Ref. [16]).
[29] H. J. Kappen and F. B. Rodríguez, in *Advances in Neural Information Processing Systems*, edited by M. S. Kearns, S. A. Solla, and D. A. Cohn (MIT Press, Cambridge, MA, 1999), Vol. 11, p. 280.
[30] G. Parisi and M. Potters, J. Phys. A **28**, 5267 (1995).
[31] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, England, 1992), p. 77.
[32] E. Marinari, G. Parisi, and F. Ritort, J. Phys. A **27**, 7647 (1994).
[33] A. J. Bray and M. A. Moore, J. Phys. C **13**, L469 (1980).
[34] R. P. Gorman and T. J. Sejnowski, Neural Networks **1**, 75 (1988).
[35] T. Tanaka, in *Advances in Neural Information Processing Systems*, edited by T. K. Leen, T. G. Diettrich, and V. Tresp (MIT Press, Cambridge, MA, 2001), Vol. 13, p. 315.
[36] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[37] J. Yedidia, in *Advanced Mean Field Methods, Theory and Practice* (Ref. [16]).
[38] J. S. Yedidia, W. T. Freeman, and Y. Weiss, in *Advances in Neural Information Processing Systems* (Ref. [35]), p. 689.
[39] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).
[40] M. Opper, in *On-Line Learning in Neural Networks*, edited by D. Saad (Cambridge University Press, Cambridge, England, 1998), p. 363.
[41] S. A. Solla and O. Winther, in *On-Line Learning in Neural Networks* (Ref. [40]), p. 379.
[42] L. Csató, E. Fokoué, M. Opper, B. Schottky, and O. Winther, in *Advances in Neural Information Processing Systems*, edited by S. A. Solla, T. K. Leen, and K.-R. Müller (MIT Press, Cambridge, MA, 2000), Vol. 12, pp. 251–257.
[43] L. Csató and M. Opper, in *Advances in Neural Information Processing Systems* (Ref. [35]), p. 444.
[44] T. P. Minka, Expectation Propagation for Approximate Bayesian Inference, Preprint MIT Media Lab, 2000 (unpublished).