Mean-field theory of Boltzmann machine learning

Toshiyuki Tanaka
Graduate School of Electrical Engineering, Tokyo Metropolitan University, 1-1 Minami Osawa, Hachioji-shi, Tokyo 192-0397 Japan
(Received 11 December 1997; revised manuscript received 27 April 1998)

I present a mean-field theory for Boltzmann machine learning, derived by employing Thouless-Anderson-Palmer free energy formalism to a full extent. Using the Plefka expansion an extended theory that takes higher-order correction to mean-field free energy formalism into consideration is presented, from which the mean-field approximation of general orders, along with the linear response correction, are derived by truncating the Plefka expansion up to desired orders. A theoretical foundation for an effective trick of using ''diagonal weights,'' introduced by Kappen and Rodríguez, is also given. Because of the finite system size and a lack of scaling assumptions on interaction coefficients, the truncated free energy formalism cannot provide an exact description in the case of Boltzmann machines. Accuracies of mean-field approximations of several orders are compared by computer simulations. [S1063-651X(98)05308-2]

PACS number(s): 84.35.+i, 05.20.−y, 75.10.Nr, 87.10.+e

I. INTRODUCTION

I present a mean-field theory for Boltzmann machine learning, which is derived by employing Thouless-Anderson-Palmer (TAP) free energy formalism [1] to a full extent. A mean-field approach to Boltzmann machine learning was suggested by Peterson and Anderson [2]. However, mean-field Boltzmann machine learning has drawn attention mainly because of its practical efficiency. Some theoretical considerations have also been made [3], but most of these are based on so-called ''naive'' mean-field theory. Recently, Kappen and Rodríguez [4] (KR hereafter) applied the linear response theorem [5] to mean-field Boltzmann machine learning. In this paper, I, extending their argument to include higher-order terms, present a mean-field theory which is fully consistent with the TAP approach in spin glass theory.

In the context of mean-field Boltzmann machine learning, the TAP approach has been mentioned in a number of studies. Galland [6] used the TAP free energy in a rather heuristic way. KR [4] also mentioned the TAP free energy, but they did not utilize it in their study. Inclusion of the Onsager reaction term was also suggested in Refs. [7, 8]. I believe this to be the first consistent treatment of the TAP formalism within the framework of mean-field Boltzmann machine learning.

The linear response theorem is also an important tool which enables us to obtain information about correlations within mean-field theory. It has been successfully applied, for example, in analyzing, within mean-field theory, a stochastic network model for correlation-based ''dynamical linking'' of features [9]. I will show that treatment of the linear response theorem within the framework based on the TAP formalism provides a quite natural and consistent argument as to how it works.

KR [4] also suggested the effective heuristics of using ''diagonal weights,'' which was justified by the fact that these gave good results. I will also provide a theoretical foundation of this ''diagonal-weight trick,'' on the basis of the framework presented in this paper.

II. BOLTZMANN MACHINE LEARNING

A Boltzmann machine with N units can be regarded as an Ising spin system having spin variables $s_i \in \{-1,1\}$, $i=1,\ldots,N$, with interactions w_{ij} between sites i and j and external fields h_i acting on sites i as its parameters. Hamiltonian $H(s)$ determining energy for each spin configuration $s=(s_1,\ldots,s_N)$ is given by

$$H(s) = -\sum_i h_i s_i - \sum_{\langle ij \rangle} w_{ij} s_i s_j , \tag{1}$$

where the notation $\langle ij \rangle$ means all distinct pairs. When values of h_i and w_{ij} are given, a Boltzmann machine represents a Boltzmann-Gibbs distribution

$$p(s) = \exp[-H(s) - \psi] = \exp\left[\sum_i h_i s_i + \sum_{\langle ij \rangle} w_{ij} s_i s_j - \psi\right], \tag{2}$$

where $-\psi$ is the Helmholtz free energy. Here, and in the sequel, I assume that the ''temperature'' is unity without loss of generality. I will identify a Boltzmann machine and the Boltzmann-Gibbs distribution represented by it, and use expressions such as ''a Boltzmann machine $p(s)$.'' For simplicity, in what follows I will argue the case without hidden units, but extension of the following argument to the case with hidden units is straightforward.

The objective of Boltzmann machine learning can be stated in terms of spin systems as follows: Determine external fields h_i and interactions w_{ij}, by knowing average magnetizations $\langle s_i \rangle_p$ and correlations $\langle s_i s_j \rangle_p$ for spins at thermal equilibrium. These averages are taken with respect to the Boltzmann-Gibbs distribution p [Eq. (2)]. This can be seen as a ''backward'' problem, and the corresponding ''forward'' problem, that is to estimate $\langle s_i \rangle_p$ and $\langle s_i s_j \rangle_p$ by knowing h_i and w_{ij}, can be solved by simulating the physical process (the Gibbs sampler [10]). Boltzmann machine learning [11] solves the backward problem by utilizing the forward problem via error feedback: Let $q(s)$ be a target

Boltzmann machine $q(s)$ with parameters h_i and w_{ij}. One knows the averages $\langle s_i \rangle_q$ and $\langle s_i s_j \rangle_q$ with respect to $q(s)$, and wants to estimate the parameters h_i and w_{ij} from these averages. Boltzmann machine learning updates current estimates of the values h_i and w_{ij} by the following learning rule:

$$\Delta h_i = \varepsilon(\langle s_i \rangle_q - \langle s_i \rangle_p), \quad \Delta w_{ij} = \varepsilon(\langle s_i s_j \rangle_q - \langle s_i s_j \rangle_p), \quad (3)$$

where the averages $\langle\ \rangle_p$ are to be evaluated from the current estimates of h_i and w_{ij} by solving the forward problem. It executes the gradient descent of Kullback divergence $D(p\|q) = \Sigma_s q(s) \ln(q(s)/p(s))$ between the target distribution $q(s)$ and the Boltzmann machine $p(s)$, whose parameters are equal to the current estimates of h_i and w_{ij}. If there are no hidden units, this learning rule provides the optimal Boltzmann machine which best approximates the target distribution [12]. If there are hidden units, it still provides a locally optimal one, but it does not assure the global optimality.

III. MEAN-FIELD THEORY

A. Exact theory

The main drawback of Boltzmann machine learning is that solving the forward problem—that is, estimating expectations by the Gibbs sampler, or exhaustively computing them—is very time consuming and hence often impractical. A mean-field theory tries to circumvent the difficulty by utilizing a mean-field approximation to solve the forward problem analytically.

In this subsection I describe an exact theory for solving the forward problem. I start with Gibbs free energy of a Boltzmann machine p with parameters h_i and w_{ij}, which is obtained by Legendre transform of Helmholtz free energy $-\psi(p)$,

$$F(p) = \left[-\psi(p) + \sum_i h_i(p) m_i(p) \right] - \sum_i h_i m_i(p). \quad (4)$$

The last term corresponds to the Zeeman energy. It should be noted that the independent variables of $F(p)$ are now $m_i(p) \equiv \langle s_i \rangle_p$ and w_{ij} by the Legendre transform, and that $h_i(p)$'s are dependent of them, whereas h_i, appearing in the Zeeman energy term should be regarded as being independent of $m_i(p)$.

Since h_i and w_{ij} are assumed to be given in the forward problem, minimization of $F(p)$ with respect to $m_i \equiv m_i(p)$ gives the true averages $m_i = \langle s_i \rangle_p$. Furthermore, $\langle s_i s_j \rangle_p$ can be obtained from $F(p)$ and the true values of $m_i = \langle s_i \rangle_p$, by using the linear response theorem [5]. This says that the susceptibility matrix

$$\chi_{ij} = \langle s_i s_j \rangle_p - \langle s_i \rangle_p \langle s_j \rangle_p \quad (5)$$

and the stability matrix

$$A_{ij} = \frac{\partial^2 F}{\partial m_i \partial m_j} \quad (6)$$

are the inverse of the others; that is, the following identity holds:

$$\sum_j \chi_{ij} A_{jk} = \delta_{ik}. \quad (7)$$

Using this theorem, one can obtain the true averages $\langle s_i s_j \rangle_p$ by first computing the stability matrix (A_{ij}) from $F(p)$, inverting it to obtain the susceptibility matrix (χ_{ij}), and then computing $\langle s_i s_j \rangle_p$ by

$$\langle s_i s_j \rangle_p = \chi_{ij} + m_i m_j \quad (i \neq j). \quad (8)$$

It should be noted that a set of the parameter values h_i and w_{ij} uniquely determines a set of the average values $\langle s_i \rangle_p$ and $\langle s_i s_j \rangle_p$ with respect to the Boltzmann-Gibbs distribution p [Eq. (2)]. In fact, there is a one-to-one correspondence between them. Therefore, the averages to be obtained by the above scenario should be the exact ones, and they provide the exact solution to the forward problem.

The difficulty of the scenario lies in the fact that one cannot write the Gibbs free energy $F(p)$ explicitly as a function of m_i, which makes the scenario intractable. Mean-field theory gives approximations of $F(p)$ as analytical functions of m_i. Using any one of the approximations, one can solve the forward problem approximately by following the scenario presented above.

B. Plefka expansion

The mean-field approximation of the Gibbs free energy has been derived in several ways. Of these, the derivation by Plefka [13] is particularly suitable for application to Boltzmann machines, since it does not regard w_{ij}'s as random quantities and hence does not require averaging over them. In spin glass theory w_{ij}'s are generally regarded as random variables, representing random interactions, and one analyzes, in the thermodynamic limit, properties which do not depend on a particular realization of w_{ij}'s. For Boltzmann machine learning, on the other hand, w_{ij}'s are given and fixed, and hence in principle they cannot be thought of as random variables.

Let us consider a Boltzmann machine p with parameters h_i and w_{ij}. In Plefka's argument, a mean-field approximation of the Gibbs free energy is systematically derived by considering the following one-parameter Hamiltonian $H(\alpha)$:

$$H(\alpha) = -\alpha \sum_{\langle ij \rangle} w_{ij} s_i s_j - \sum_i h_i s_i, \quad (9)$$

and then expanding the true Gibbs free energy $F(\alpha)$ for the Hamiltonian into the power series of α:

$$F(\alpha) = F(0) + \alpha F'(0) + \tfrac{1}{2} \alpha^2 F''(0) + \tfrac{1}{6} \alpha^3 F'''(0) + \cdots, \quad (10)$$

where $F'(\alpha) \equiv \partial F/\partial \alpha$, $F''(\alpha) \equiv \partial^2 F/\partial \alpha^2$, and so on. This expansion is called the Plefka expansion. Note that the derivatives with respect to α should be taken with m_i fixed. Since $H(\alpha = 1)$ is the original Hamiltonian (1) to be considered, setting $\alpha = 1$ in Eq. (10) yields, leaving the convergence problem aside, the true Gibbs free energy, i.e., $F(p) \equiv F(1)$.

The coefficients of the Plefka expansion up to third order are given as follows [14].

$$F(0)=\frac{1}{2}\sum_i\left[(1+m_i)\ln\left(\frac{1+m_i}{2}\right)+(1-m_i)\ln\left(\frac{1-m_i}{2}\right)\right]-\sum_i h_i m_i, \tag{11}$$

$$F'(0)=-\sum_{\langle ij\rangle} w_{ij}m_i m_j, \tag{12}$$

$$F''(0)=-\sum_{\langle ij\rangle} w_{ij}^2(1-m_i^2)(1-m_j^2), \tag{13}$$

$$F'''(0)=-4\sum_{\langle ij\rangle} w_{ij}^3 m_i m_j(1-m_i^2)(1-m_j^2)-6\sum_{\langle ijk\rangle} w_{ij}w_{jk}w_{ik}(1-m_i^2)(1-m_j^2)(1-m_k^2). \tag{14}$$

In the above, $\langle ijk\rangle$ means that the summation should be taken over all distinct triplets. By truncating the Plefka expansion up to the nth-order term and letting $\alpha=1$, one can obtain the nth-order approximation F_n of the Gibbs free energy F. Note that F_1 is identical to the Weiss free energy, and F_2 is the TAP free energy for Sherrington-Kirkpatrick (SK) models: For the case of Boltzmann machines with finite N, one cannot expect in general that higher-order terms vanish, and F_n is indeed an approximation of F.

C. Mean-field approximation

1. Forward problem

From the nth-order approximation of the Gibbs free energy, one can construct the nth-order method of mean-field approximation. In this subsection I describe the methods of several orders for the forward problems. In Sec. III C 2 I will discuss the methods for the backward problems.

As described above, minimization of the Gibbs free energy F with respect to m_i gives the true value of m_i. Using the nth-order approximation F_n in place of F, one can obtain an nth-order estimate of m_i by minimizing F_n with respect to m_i. This minimization problem can be solved by considering the stationary conditions $\partial F_n/\partial m_i=0$, $i=1,\ldots,N$. These conditions constitute the self-consistent equations of the nth-order mean-field approximation. For example, when $n=1$, the conditions are

$$\tanh^{-1} m_i-h_i-\sum_{j\neq i} w_{ij}m_j=0, \tag{15}$$

which are those of Weiss mean-field theory. For $n=2$ these give the TAP equations for SK models:

$$\tanh^{-1} m_i-h_i-\sum_{j\neq i} w_{ij}m_j+\sum_{j\neq i} w_{ij}^2(1-m_j^2)m_i=0. \tag{16}$$

Self-consistent equations for still higher orders can be obtained in the same way. Because F_n is an approximation of F, solutions m_i of the stationary conditions are not exact. Moreover, they are not necessarily unique, as extensively studied in spin glass literature [15].

The linear response theorem provides a practical basis for the linear response *correction* [4], which gives an approximate estimate of the correlations $\langle s_i s_j\rangle_p$ in the mean-field approximation. Using the solution $\{m_i\}$ of the self-consistent equations, the approximated stability matrix

$$A_{ij}^{(n)}=\frac{\partial^2 F_n}{\partial m_i \partial m_j} \tag{17}$$

is evaluated in terms of the already known quantities m_i and w_{ij}. For example, $A_{ij}^{(1)}$ is given by

$$A_{ij}^{(1)}=\frac{1}{1-m_i^2}\delta_{ij}-w_{ij}. \tag{18}$$

w_{ij} for $i=j$ is undefined, and should be regarded as 0 at this point. Although the linear response theorem no longer holds exactly since $(A_{ij}^{(n)})$ is not exact, one can expect that it still holds approximately. Thus inverting $(A_{ij}^{(n)})$ yields the approximated susceptibility matrix $(\chi_{ij}^{(n)})$. Then, using relation (5), while substituting $\chi_{ij}^{(n)}$ and m_i in place of χ_{ij} and $\langle s_i\rangle_p$, respectively, one can compute an nth-order estimate of $\langle s_i s_j\rangle_p$. This constitutes the linear response correction in the nth-order mean-field approximation.

So far, under the condition that h_i and w_{ij} of a Boltzmann machine p are all known, one can estimate $\langle s_i\rangle_p$ and $\langle s_i s_j\rangle_p$ from the self-consistent equations and the linear response correction of nth-order mean-field approximation, respectively, in the way described above, which defines a method of solving the forward problem approximately. I will call the method the nth-order method for the forward problem.

2. Backward problem

For Boltzmann machines without hidden units, one can solve the backward problem directly by using the mean-field approximation, *without* referring to the error feedback scheme (3) employed in the ordinary Boltzmann machine learning. Assume that $\langle s_i\rangle_p$ and $\langle s_i s_j\rangle_p$ are all known. Then the susceptibility matrix (χ_{ij}) is computed exactly using Eq. (5). Inverting it yields the (exact) stability matrix (A_{ij}). On the other hand, the stability matrix has under the nth-order mean-field approximation an analytic expression derived from Eq. (17), which is a function of $m_i=\langle s_i\rangle_p$ and w_{ij}. Equating them gives the simultaneous equations from which w_{ij}'s are to be determined, since $m_i=\langle s_i\rangle_p$ are assumed to be known. Once w_{ij}'s are obtained, one can evaluate h_i using the self-consistent equations of the nth-order approximation. The procedure described above constitutes the nth-order method for solving the backward problem.

It should be noted that, although evaluation of h_i's is an appropriately defined problem—since the number of variables h_i to be determined, N, is equal to the number of the self-consistent equations—evaluation of w_{ij}'s bears a problem of overdetermination because the number of variables w_{ij} to be determined is $N(N-1)/2$, whereas the number of independent equalities of Eq. (7) is $N(N+1)/2$ including those for the diagonals, since the susceptibility and stability matrices are symmetric. These equations should be simulta-

neously satisfied by the true w_{ij}'s in the *exact* theory, where m_i's are also true, but in *approximate* theory it is no longer expected for them to hold simultaneously. There may be several choices for dealing with the overdetermination, both for the forward and backward problems. KR [4] employed a heuristics of using the $N(N-1)/2$ equations on the off-diagonals of the *susceptibility* matrix for both of the problems. The heuristics of KR has a merit of consistency in that, given a Boltzmann machine p, when one solves the backward problem using a solution of the forward problem with p one obtains a result which is identical with p. Even though the consistency property is the preferable one, an argument of how these choices affect the degree of accuracy of the approximation will be very intricate. Therefore, I leave the problem of how to cope with the overdetermination beyond the scope of this paper.

IV. DIAGONAL-WEIGHT TRICK

Within the framework presented in this paper, the KR method can be regarded as a variant of the first-order method, since it is based on the Weiss free energy F_1. Numerical experiments [4] reveal, however, that it performs far better than the first-order method described above. The difference between KR and first-order methods is that KR suggested using ''diagonal weights'' w_{ii}, which are to be defined by

$$A_{ii}=\frac{1}{1-m_i^2}-w_{ii}. \tag{19}$$

This equation can be regarded as a naive application of Eq. (18) to the diagonals $i=j$, without regarding w_{ii} to be zero. The self-consistent equation (15) is rewritten accordingly as

$$\tanh^{-1} m_i-h_i-\sum_j w_{ij}m_j=0, \tag{20}$$

where the summation now includes the diagonal term $w_{ii}m_i$. An obvious merit of this ''diagonal-weight trick'' is that it practically resolves the overdetermination problem discussed in Sec. III C 2, because the number of variables to be determined is now $N(N+1)/2$, which is equal to the number of independent equations. Thus one can simply take the $N(N+1)/2$ equations for the elements of the stability matrix in order to determine $N(N+1)/2$ variables w_{ij}. The most important advantage of the diagonal-weight trick is, however, that it is what enables the KR method to perform substantially better than the first-order method, as shown in this section.

I present an explanation of the diagonal-weight trick on the basis of the third-order theory. In the third-order approximation the diagonal weights w_{ii} determined by the trick (19) are given by

$$\begin{aligned} w_{ii}=&-\sum_{j\neq i}(w_{ij})^2(1-m_j^2)\\ &-4\sum_{j\neq i}(w_{ij})^3m_im_j(1-m_j^2)\\ &-2\sum_{\langle i|jk\rangle}w_{ij}w_{ik}w_{jk}(1-m_j^2)(1-m_k^2). \end{aligned} \tag{21}$$

The notation $\langle i|jk\rangle$ means that the summation should be taken over all distinct triplets with i fixed. The first term comes from the second-order term of the Gibbs free energy, and the remaining two terms from the third-order term. It should be noted that the $w_{ij}w_{ik}w_{jk}$ terms are expected to be dominant among the third-order terms, since the number of such terms is $O(N^2)$, whereas the number of $(w_{ij})^3$ terms is $O(N)$. The diagonal term $w_{ii}m_i$ can be written as

$$\begin{aligned} w_{ii}m_i=&-\frac{\partial}{\partial m_i}\left(\frac{1}{2}F''(0)+\frac{1}{6}F'''(0)\right)\\ &-\frac{2}{3}\sum_{j\neq i}(w_{ij})^3(1+3m_i^2)m_j(1-m_j^2). \end{aligned} \tag{22}$$

Therefore, the self-consistent equation (20) with the diagonal term eventually takes into account the second-order term $\frac{1}{2}(\partial F''(0)/\partial m_i)$ as a whole, and the dominant part of the third-order term $\frac{1}{6}(\partial F'''(0)/\partial m_i)$:

$$\begin{aligned} &\tanh^{-1} m_i-h_i-\sum_j w_{ij}m_j\\ &=\frac{\partial}{\partial m_i}F_1-w_{ii}m_i\\ &=\frac{\partial}{\partial m_i}F_3+\frac{2}{3}\sum_{j\neq i}(w_{ij})^3(1+3m_i^2)m_j(1-m_j^2). \end{aligned} \tag{23}$$

This shows that, if the second-order approximation of the free energy is exact, the first-order method with the diagonal-weight trick and the second-order method without it will give the same results. Even if the third-order term of the free energy cannot be neglected, the dominant part is incorporated by the diagonal term, and one can expect that the result will not be so different from the result of the third-order method without the trick, if the part incorporated by the diagonal term is indeed dominant.

It is possible to extend the argument to still higher orders. In the nth-order coefficient $F^{(n)}(0)$, $n\geq 3$, of the Plefka expansion (10), the part which is dominant in the number of terms $[O(N^n)]$ is given by [14]

$$-n!\sum_{\langle i_1,i_2,\ldots,i_n\rangle}w_{i_1i_2}w_{i_2i_3}\cdots w_{i_ni_1}(1-m_{i_1}^2)(1-m_{i_2}^2)\cdots(1-m_{i_n}^2). \tag{24}$$

The dominant part contributes to the self-consistent equations by the terms

$$\frac{1}{n!}\frac{\partial F^{(n)}(0)}{\partial m_i} \approx 2 \sum_{\langle i|i_2,\ldots,i_n\rangle} w_{ii_2} w_{i_2 i_3} \cdots w_{i_n i} (1-m_{i_2}^2)\cdots(1-m_{i_n}^2) m_i , \tag{25}$$

and to the diagonal weight w_{ii} by

$$-\frac{1}{n!}\frac{\partial^2 F^{(n)}(0)}{\partial m_i^2} \approx -2 \sum_{\langle i|i_2,\ldots,i_n\rangle} w_{ii_2} w_{i_2 i_3} \cdots w_{i_n i} (1-m_{i_2}^2)\cdots(1-m_{i_n}^2). \tag{26}$$

It is evident in these expressions that the latter, as a component of the diagonal weight, exactly cancels out the former in the self-consistent equation. Therefore, it has been shown that the diagonal term $w_{ii}m_i$ in the self-consistent equation effectively cancels out the dominant parts coming from the nth-order terms $(n \geqslant 2)$ of the Plefka expansion.

The above discussion shows that the KR method exhibits a performance superior to the conventional first-order method because it effectively incorporates higher-order contributions via the diagonal-weight trick. It is possible to combine the trick with the second- or third-order method, by eliminating from the self-consistent equations the terms already incorporated by the diagonal term. The self-consistent equation of the second-order method with the diagonal-weight trick is the same as Eq. (20), that of the first-order method with the trick, because the entire second-order contribution to the self-consistent equation has already been incorporated by the diagonal term. Similarly, the self-consistent equation of the third-order method with the trick is given explicitly by the simple formula

$$\tanh^{-1} m_i - h_i - \sum_j w_{ij} m_j (1+c_{ij}) = 0, \tag{27}$$

where $c_{ii}=0$,

$$c_{ij} = \tfrac{2}{3}(w_{ij})^2 (1+3m_i^2)(1-m_j^2) \tag{28}$$

for $i \neq j$, and w_{ii} is determined by Eq. (19).

Use of the diagonal-weight trick is especially advantageous for solving the backward problems, because it saves the amount of computation to a considerable extent. For the forward problems, however, the diagonal weights w_{ii} should be determined so that the diagonal terms $\chi_{ii}^{(n)}$ of the approximated susceptibility matrix are equal to $1-m_i^2$, which would require iterative computation and does not seem practical.

V. COMPARISON OF PERFORMANCE BY COMPUTER SIMULATIONS

Thus far, a family of mean-field-theory-based methods of various orders has been obtained systematically from the Plefka expansion. It is then important to compare the performance of these methods. Since the Plefka expansion is essentially a Taylor expansion, the accuracy depends on the order of expansion as well as the magnitude of w_{ij}'s in a quite complicated manner. One can expect, for sufficiently small w_{ij}'s, that higher-order methods give better results. However, when w_{ij}'s are large, higher-order methods may give erroneous results. I compared the accuracy of the methods for various orders using computer simulations, not to propose novel efficient algorithms, but to gain insight into how each of the methods will perform under certain conditions.

We restrict ourselves to cases where there are no hidden units. As already mentioned, in such cases one can directly solve the ''backward'' problem of estimating the parameters h_i and w_{ij} from the expectations $\langle s_i \rangle_p$ and $\langle s_i s_j \rangle_p$, which are assumed to be observed exactly. We also focus on the cases where the number of units N is small, in order to evaluate the accuracy of results explicitly by means of Kullback divergence. In the simulations the target distribution $p(s)$, defining $\langle s_i \rangle_p$ and $\langle s_i s_j \rangle_p$, was assumed to be given by another Boltzmann machine, whose parameters h_i and w_{ij} were chosen as independent random variables following a Gaussian distribution with mean $\mu=0$ and variance σ^2.

A. Description of algorithms

In this subsection I summarize the actual procedures implementing the methods used in the simulations. Since the simulations treat the backward problems, the algorithms basically follow the description in Sec. III C 2. An important issue is the way to resolve the overdetermination. In this paper I employ the simplest heuristics to use the $N(N-1)/2$ equations on the off-diagonal elements of the *stability* matrix:

$$A_{ij} = \frac{\partial^2 F_n}{\partial m_i \partial m_j} \quad (i \neq j). \tag{29}$$

The algorithms of the methods without the diagonal-weight trick are summarized as follows:

(1) Compute the susceptibility matrix (χ_{ij}) from the observed $\langle s_i \rangle_p$ and $\langle s_i s_j \rangle_p$ using Eq. (5).

(2) Invert it to obtain the (exact) stability matrix (A_{ij}).

(3) Compute w_{ij}'s by solving Eq. (29) with $m_i = \langle s_i \rangle_p$.

(4) Compute h_i's by substituting m_i and w_{ij} into the self-consistent equations.

Equation (29) reads, for the first-order method,

$$A_{ij} = -w_{ij}; \tag{30}$$

for the second-order method,

$$A_{ij} = -w_{ij} - 2(w_{ij})^2 m_i m_j; \tag{31}$$

and, for the third-order method,

$$A_{ij} = -w_{ij} - 2(w_{ij})^2 m_i m_j$$

$$-\frac{2}{3}(w_{ij})^3(1-3m_i^2)(1-3m_j^2)$$

$$-4\sum_{k\neq i,j} w_{ij}w_{ik}w_{jk}m_im_j(1-m_k^2), \qquad (32)$$

all for $i\neq j$. For the first- and second-order methods, Eqs. (30) or (31) can be solved independently for each w_{ij}. For the third-order method, however, Eq. (32) should be solved simultaneously. In the following simulations this was done by the gradient-descent method on the squared residual error. Another practical difficulty is that Eqs. (31) and (32) are nonlinear, and may have more than one solution. Considering the origin of these equations that have been obtained from Taylor expansion, one can expect continuity of the solution as higher-order effects are gradually increased from zero, provided that the approximation based on the expansion is valid. Based on the continuity property one can pick up the relevant solution. In the following simulations, for the second-order method this was done analytically. For the third-order method it was done numerically, pursuing the solution while increasing the magnitude of the third-order terms. In fact this approach requires a considerable amount of computation and hence is less practical, but it is expected to retain the reliability of the obtained solution. Of course, in view of the amount of computation there may be other approaches.

Self-consistent equations for the first- and second-order methods are given by Eqs. (15) and (16), respectively. The self-consistent equation for the third-order method is

$$\tanh^{-1} m_i - h_i - \sum_{j\neq i} w_{ij}m_j + \sum_{j\neq i} w_{ij}^2(1-m_j^2)m_i$$

$$-\frac{2}{3}\sum_{j\neq i}(w_{ij})^3(1-3m_i^2)m_j(1-m_j^2)$$

$$+2\sum_{\langle i|jk\rangle} w_{ij}w_{ik}w_{jk}m_i(1-m_j^2)(1-m_k^2)=0. \qquad (33)$$

Using only the zeroth-order term of the Plefka expansion, along with the Zeeman energy term, gives a method which I call the zeroth-order method. It is defined as

$$h_i=\tanh^{-1} m_i, \quad w_{ij}=0. \qquad (34)$$

The algorithms of the methods with the diagonal-weight trick are defined as follows:

(1) Compute the susceptibility matrix (χ_{ij}) from the observed $\langle s_i\rangle_p$ and $\langle s_is_j\rangle_p$ using Eq. (5).

(2) Invert it to obtain the (exact) stability matrix (A_{ij}).

(3) Compute w_{ij}'s by solving Eq. (29) for $i\neq j$, and Eq. (19) for $i=j$, with $m_i=\langle s_i\rangle_p$.

(4) Compute h_i's by substituting m_i and w_{ij} into the self-consistent equations.

The self-consistent equations for the first- and second-order methods are both given by Eq. (20). For the third-order method I used Eq. (27) to supplement the terms which are not taken into account by the diagonal term [see Eq. (23)].

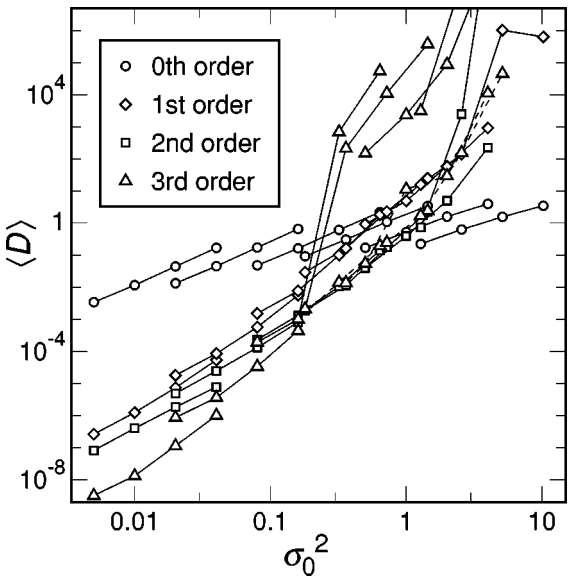

FIG. 1. Average Kullback divergence $\langle D\rangle$ between the randomly generated targets and the Boltzmann machines determined by a mean-field approximation of several orders without the diagonal-weight trick, vs $\sigma_0^2\equiv(N-1)\sigma^2$. Results for $N=3$, 5, 9, and 17 and $\sigma=0.05$, 0.1, 0.2, 0.3, 0.5, and 0.8 are shown. The full lines connecting markers show that the markers are the results for the same value of σ. The dashed lines indicate that the markers are the results of the third-order method when the averages are taken over ''reliable'' trials only.

The approaches taken to resolve the overdetermination and the multiple solutions are the same as those employed in the methods without the diagonal-weight trick.

B. Results and discussion

1. Methods without diagonal-weight trick

In the first set of simulations I compared the accuracy of the zeroth-, first-, second-, and third-order methods without the diagonal-weight trick. The Kullback divergence D between the target and the result given by each of the methods was measured, and the averages $\langle D\rangle$ of 200 trials were plotted in Fig. 1. The horizontal axis is scaled with the normalized variance $\sigma_0^2\equiv(N-1)\sigma^2$.

For small σ_0, higher-order methods had consistently lower average divergences, and thus the third-order method gave the best results among all the methods investigated. This indicates that the inclusion of higher-order terms certainly improved the accuracy in such cases. On the other hand, for larger σ_0 the third-order method became numerically rather unstable and sometimes gave results with erroneously large divergences, which make the averaged divergence $\langle D\rangle$ large. For still larger σ_0 the second- and first-order methods also showed similar behaviors. The erroneous behavior is not hard to detect for each trial since it appears rather drastically. I investigated this behavior in more detail for the third-order method, since it showed the behavior most evidently. For measuring to what extent a result given by the third-order method is unreliable, I used the ratio D_3/D_2 of the two divergences $D_3=D(p_{3\mathrm{rd}}\|p)$ and $D_2=D(p_{2\mathrm{nd}}\|p)$ for a target p, the former being the one between the target p and the result $p_{3\mathrm{rd}}$ given by the third-order method, and the latter

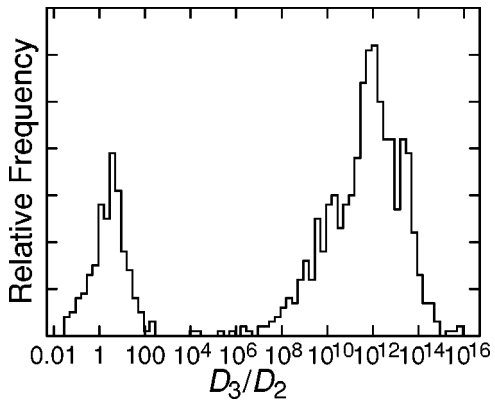

FIG. 2. Histogram of the ratio D_3/D_2 for the cases $N=12$ and $\sigma=0.4$. Results of 1000 trials are shown.

between the target p and the result $p_{2\mathrm{nd}}$ given by the second-order method. As an example, Fig. 2 shows a histogram of the ratio D_3/D_2 for the case where $N=12$ and $\sigma=0.4$. In the following, a heuristic criterion, that trials with a ratio D_3/D_2 greater than 100 are judged as unreliable, was used throughout. Shown in Fig. 1 by dashed lines are the divergences averaged over trials which were judged as reliable with the criterion above.

This instability of the higher-order methods can be explained in terms of the phase transition in spin glass theory. The condition for phase transition into the spin glass phase can be expressed as $\sigma_0^2=1$ for the SK model ($h_i\equiv 0$). When σ_0^2 is larger than this point, it is known that the self-consistent equation for the SK model develops many solutions. In the spin glass phase the free energy has a so-called many-valley structure, and each valley is separated from other valleys by barriers of infinite height. Some of the solutions of the self-consistent equation are expected to correspond to such valleys, and when the state is trapped within one of the valleys the time average of spin variables will be given by one such solution. Intuitively speaking, the phase transition phenomenon, or bifurcation of solutions, also occurs in Boltzmann machines, and results in the observed erroneous behavior of higher-order methods. Figure 3 shows a result which supports this explanation. It shows how the proportion of reliable trials depends on σ_0^2 for the third-order method. The proportion decreases from 1 to 0 around σ_0^2 ~ 1 as expected by spin glass theory, and it is observed that the slope becomes steeper as N increases, suggesting that it reflects the phase transition phenomenon.

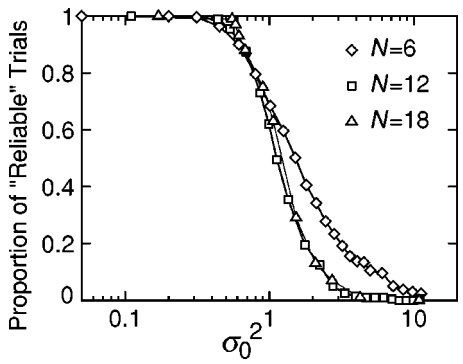

FIG. 3. Proportion of ''reliable'' trials to total trials for the third-order method. The number of total trials are 500, 200, and 100 for $N=6$, 12, and 18, respectively.

It can be stated that the effectiveness of the methods based on mean-field theory for Boltzmann machine learning is affected by the phase transition phenomenon. It should be noted that, as for Boltzmann machine learning, the relevant quantities are not the time averages but the Gibbs averages. Once all the parameters h_i and w_{ij} are given, the Gibbs averages are in principle completely determined by these parameters, and thus the correct solution for $\{m_i\}$ should be unique. Therefore, when the self-consistent equation has many solutions, elaborating any one of these solutions is of use for evaluating the time averages, but will be of little use for evaluating the Gibbs averages, which are the relevant quantities. The methods based on mean-field theory work well in the ''paramagnetic'' region (i.e., all w_{ij} 's are relatively small and the solution of the self-consistent equations is unique) and higher-order methods give more accurate results, whereas, in the ''spin glass'' region beyond the phase transition point (w_{ij}'s are large), methods based on mean-field theory can provide many spurious results. In such conditions even the first-order method becomes rather unstable and gives erroneous results, while the simple zeroth-order method seems to be robust against the instability, as indicated by the large σ_0^2 part of Fig. 1. It should be noted, from a practical viewpoint, that the ''spin glass'' phase limits the usefulness not only of mean-field-theory-based methods but also of the direct method using the Gibbs sampler to solve the ''forward'' problem. The Gibbs sampler evaluates the time averages for estimating the Gibbs averages by simulating the physical process, and it will, as is the case with real spin glasses, become trapped in one of the valleys for a considerably long time, which will prevent it from correctly estimating the Gibbs averages from the time averages.

2. Methods with diagonal-weight trick

In order to investigate how the diagonal-weight trick affects accuracy, I executed another set of simulations with methods employing the diagonal-weight trick. The sets of targets used in these simulations were the same as those in previously described simulations without the trick. The results, obtained as average Kullback divergences $\langle D\rangle$, are summarized in Fig. 4.

When σ_0^2 is small, the third-order method with the diagonal-weight trick marks the best performance on average. The first- and second-order methods showed almost the same performance, and thus these two methods are hard to distinguish on the figure. The performance of these two methods was also almost the same as that of the second-order method *without* the diagonal-weight trick, as shown in Fig. 5. These observations show that the first-order method with the diagonal-weight trick performs as good as the second-order methods owing to the trick, which means that the trick is quite effective in such conditions. It is also consistent, up to second order, with the analytical argument on the trick presented in Sec. IV. The self-consistent equations for these three methods [Eqs. (16) and (20)] are effectively identical up to the second order. The second-order term is identical to the Onsager reaction term of SK models. As is well known in spin glass theory, this cannot be neglected in considering

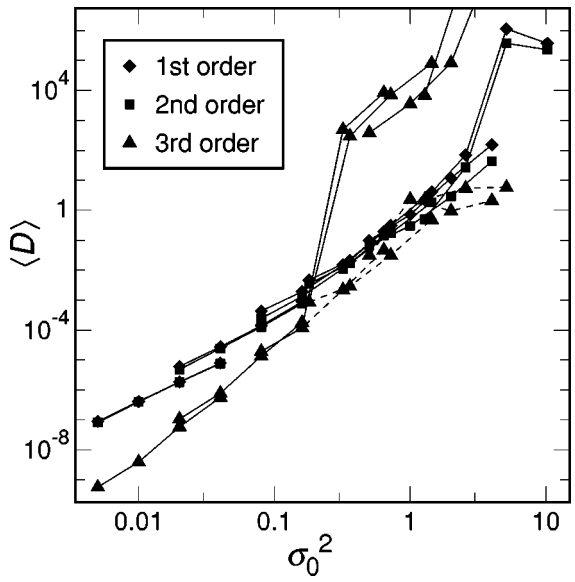

FIG. 4. Average Kullback divergence of methods of several orders with the diagonal-weight trick. Results for $N=3$, 5, 9, and 17 and $\sigma=0.05$, 0.1, 0.2, 0.3, 0.5, and 0.8 are shown.

such randomly frustrated systems, and its inclusion in the simulations certainly improved the accuracy, as expected. The equations giving the linear response correction are different for these methods [Eqs. (30) and (31)] in the second-order term $-2(w_{ij})^2 m_i m_j$, but this seems to have little effect on the results in this case.

One may suppose that the second-order term $-2(w_{ij})^2 m_i m_j$ has little effect when σ_0^2 is small because in such cases m_i's are close to zero, and therefore this second-order term becomes vanishingly small compared with w_{ij}. To investigate this, still another set of simulations was executed, in which the parameters h_i of the target distribution $p(s)$ were chosen as independent Gaussian random variables with variance σ^2 and mean $\mu=0.8$ instead of 0. This setup allows us to check the effect of m_i's, which are now no longer small, while w_{ij}'s remain small as before. The results are summarized in Fig. 6. The first-order method with the trick showed a slightly better performance than the first-order method without the trick, implying that the incorporation of the diagonal weight has an effect of improving accuracy again in this case. However, its performance was considerably worse than the second-order methods. It shows that the second-order term $-2(w_{ij})^2 m_i m_j$ certainly affects the accuracy when h_i's (hence m_i's) are not close to zero.

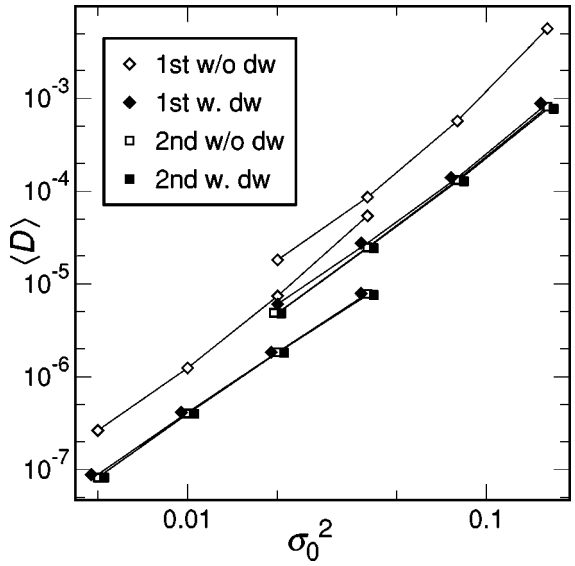

FIG. 5. Average Kullback divergence of the first- and second-order methods with and without the diagonal-weight trick. Results for $N=3$, 5, 9, and 17 and $\sigma=0.05$ and 0.1 are shown.

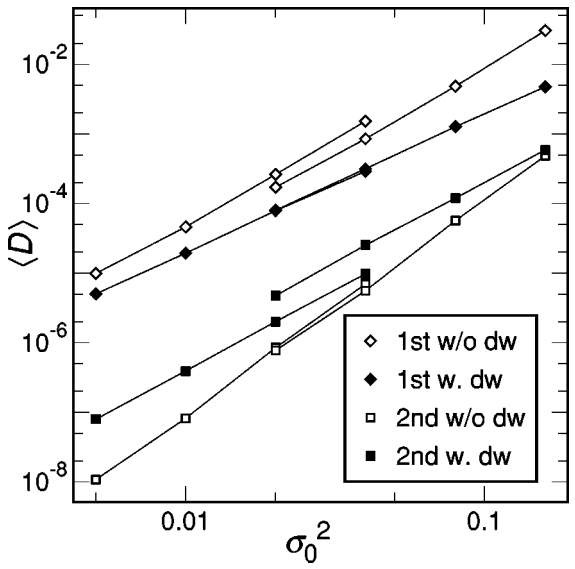

FIG. 6. Average Kullback divergence of the first- and second-order methods with and without the diagonal-weight trick. The mean μ of the parameters h_i is set to 0.8. Results for $N=3$, 5, 9, and 17 and $\sigma=0.05$ and 0.1 are shown.

When considering up to the third order, it seems that these results do not correspond to the analytical argument. According to the argument, the effective incorporation of ''dominant parts'' of the third-order contribution via the diagonal term, if it works as anticipated, will have the second-order method with the diagonal-weight trick outperform the second-order method without the trick, and lie closer to the third-order method without the trick. The results in Figs. 1 and 5 showed that the former two methods were almost the same in performance, whereas the latter outperformed these two. This means that the effective incorporation of dominant parts by the diagonal term did not work efficiently in those situations investigated in these simulations, which would arguably be due to small N. For the case where h_i's are biased with $\mu=0.8$, the second-order method without the trick showed better performance than the second-order method with the trick when σ_0^2 is small. In fact, the performance of the former was almost equal to the performance of the third-order method without the trick. From these results it should be stated that the effectiveness of the diagonal-weight trick varies between cases. As far as I investigated, it was observed as a general trend that the trick improved accuracy with h_i's smaller compared with w_{ij}'s, whereas it gave worse results than without it when h_i's were larger.

When σ_0^2 becomes large, instability appears just as it does for methods without the diagonal-weight trick. It was observed that higher-order methods are more sensitive, and that employing the diagonal-weight trick makes them somewhat

less sensitive than without it, but these tendencies are not so apparent in the figure. In the region where $\sigma_0^2 \gg 1$ the zeroth-order method was the best among all the methods investigated. This again supports the argument presented above that, in such cases, higher-order treatments intended to give more precision are actually not effective and give erroneous results, and that the simplest zeroth-order method seems to be robust against the instability.

VI. CONCLUSION

I have presented a unified theory of the mean-field Boltzmann machine learning based on TAP free energy formalism and incorporating the linear response theorem in a consistent way. By utilizing the Plefka expansion, an extended theory including higher-order terms has been discussed, from which a mean-field approximation of general orders is derived by truncating the Plefka expansion up to desired orders. Based on the framework, a theoretical foundation of an effective trick of using diagonal weights, introduced by Kappen and Rodríguez, has been also given.

Computer simulations for comparing the accuracy of the methods of various orders have shown that higher-order methods give better results when the parameters h_i and w_{ij} of the target are small. On the other hand, when the parameters are large, higher-order methods exhibit instability and may give erroneous results, while the simplest zeroth-order method is robust against such instability and gives the best results. These observations have been explained in terms of the spin glass phase transition, and this seems to limit the effectiveness of the methods based on mean-field theory for Boltzmann machine learning, as well as the direct method using the Gibbs sampler.

When the parameters are small, the first-order method with the diagonal-weight trick has shown a performance as good as the second-order methods, which supports the effectiveness of the KR method and the trick in such conditions. However, the effectiveness of the trick in general varies between cases, as was shown by computer simulations.

In this paper we have left aside the issue of computational complexity. When w_{ij}'s are small the higher-order methods give better results, but at the expense of increasing computation time. As mentioned by KR [4], the computation time required for the KR method is $O(N^3)$ because it includes inversion of a matrix of size N. This is the same order as that of the second-order method. However, the third-order method is computationally much heavier because it has to incorporate an iterative calculation for solving nonlinear simultaneous equations. For practical applications it seems that appropriate algorithms and/or approximation schemes are necessary, and we expect that the theory presented in this paper will be useful in deriving such schemes.

ACKNOWLEDGMENTS

I would like to thank Dr. Masato Okada for drawing my attention to the work by Kappen and Rodríguez, and Dr. Kazuo Nakanishi for pointing out an error in a previous version of the paper.

[1] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philos. Mag. **35**, 593 (1977).
[2] C. Peterson and J. R. Anderson, Complex Syst. **1**, 995 (1987).
[3] G. E. Hinton, Neural Comput. **1**, 143 (1989).
[4] H. J. Kappen and F. B. Rodríguez, Neural Comput. **10**, 1137 (1998).
[5] G. Parisi, *Statistical Field Theory* (Addison-Wesley, Reading, MA, 1988).
[6] C. C. Galland, Network **4**, 355 (1993).
[7] A. P. Dunmur and D. M. Titterington, in *Advances in Neural Information Processing Systems*, edited by M. C. Mozer, M. I. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1997), Vol. 9, pp. 431–437.
[8] T. Hofmann and J. M. Buhmann, IEEE Trans. Pattern Anal. Mach. Intell. **19**, 1 (1997); **19**, 197(E) (1997).
[9] H. J. Kappen, Phys. Rev. E **55**, 5849 (1997).
[10] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 721 (1984).
[11] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cogn. Sci. **9**, 147 (1985).
[12] S. Amari, K. Kurata, and H. Nagaoka, IEEE Trans. Neural Netw. **3**, 260 (1992).
[13] T. Plefka, J. Phys. A **15**, 1971 (1982).
[14] K. Nakanishi and H. Takayama, J. Phys. A **30**, 8085 (1997).
[15] K. Nemoto, J. Phys. A **21**, L287 (1988).