Empirical Bayes method for Boltzmann machines

OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **53** (2020) 014004 (20pp)
https://doi.org/10.1088/1751-8121/ab57a7

Empirical Bayes method for Boltzmann machines

Muneki Yasuda[1] **and Tomoyuki Obuchi**[2]

[1] Graduate School of Science and Engineering, Yamagata University, Yonezawa, Yamagata 992-8510, Japan
[2] Department of Mathematical & Computing Science, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152-8552, Japan

E-mail: muneki@yz.yamagata-u.ac.jp

Received 17 June 2019, revised 21 October 2019
Accepted for publication 14 November 2019
Published 10 December 2019

Abstract

We consider an empirical Bayes method for Boltzmann machines and propose an algorithm for it. The empirical Bayes method allows for estimation of the values of the hyperparameters of the Boltzmann machine by maximizing a specific likelihood function referred to as the empirical Bayes likelihood function in this study. However, the maximization is computationally hard because the empirical Bayes likelihood function involves intractable integrations of the partition function. The proposed algorithm avoids this computational problem by using the replica method and the Plefka expansion. Our method is quite simple and fast because it does not require any iterative procedures and gives reasonable estimates at a certain condition. However, our method introduces a bias to the estimate, which exhibits an unnatural behavior with respect to the size of the dataset. This peculiar behavior is supposed to be due to the approximate treatment by the Plefka expansion. A possible extension to overcome this behavior is also discussed.

Keywords: Boltzmann machine, inverse Ising problem, empirical Bayes method, replica method, Plefka expansion

1. Introduction

Boltzmann machine learning (BML) [1] has been actively studied in the field of machine learning and also in statistical mechanics. In statistical mechanics, the problem of BML is sometimes referred to as the *inverse Ising problem*, because a Boltzmann machine is the same

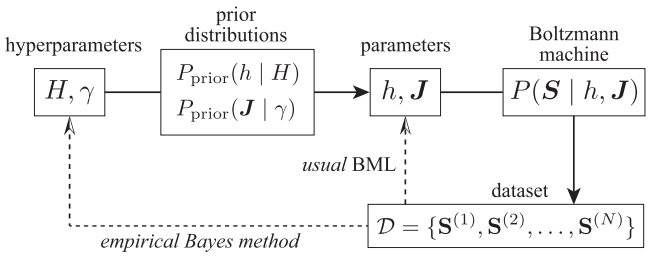

Figure 1. Illustration of scheme of empirical Bayes method considered in this study.

as an Ising model, and BML can be regarded as an inverse problem for the Ising model. The framework of the *usual* BML is as follows. Given a set of observed data points (e.g. spin snapshots), we estimate appropriate values of the parameters, the external field and couplings, of our Boltzmann machine through maximum likelihood (ML) estimation (see section 2.1). Because BML involves intractable multiple summations (i.e. evaluation of the partition function), many approximations for it were proposed from the viewpoint of statistical mechanics [2]: for example, methods based on mean-field approximations (such as the Plefka expansion [3] and the cluster variation method [4]) [5–11] and methods based on other approximations [12, 13].

In this study, we focus on another type of learning problem. We consider prior distributions of parameters of the Boltzmann machine and assume that the prior distributions are governed by some hyperparameters. The introduction of the prior distributions is strongly connected with the regularized ML estimation, in which the hyperparameters can be regarded as regularization coefficients (see section 2.1). The regularized ML estimation is important in preventing over-fitting to the dataset. In particular, the over-fitting problem becomes serious for small datasets. As mentioned above, the aim of the *usual* BML is to optimize the values of the parameters of the Boltzmann machine by using a set of observed data points. Meanwhile, the aim of the problem investigated in this study is the estimation of appropriate values of the hyperparameters from the dataset without estimating specific values of the parameters. One way to allow us to accomplish this from the Bayesian point of view is the *empirical Bayes method* (or also called type-II ML estimation or evidence approximation) [14, 15] (see section 2.2). The schemes of the *usual* BML and of our problem are illustrated in figure 1.

However, the evaluation of the likelihood function in the empirical Bayes method is again intractable, because it involves intractable multiple integrations of the partition function. In this study, we analyze the empirical Bayes method for fully-connected Boltzmann machines, using statistical mechanical techniques based on the replica method [16, 17] and the Plefka expansion to derive an algorithm for it. We consider two types of cases of the prior distribution of $\boldsymbol{J}$: the cases of Gaussian and Laplace priors.

The rest of this paper is organized as follows. The formulations of the *usual* BML and the empirical Bayes method are presented in section 2. In section 3, we describe our statistical mechanical analysis for the empirical Bayes method. The proposed inference algorithm obtained from our analysis is shown in section 3.3 with its pseudocode. In section 4, we examine our proposed method through numerical experiments. Finally, the summary and some discussions are presented in section 5.

2. Boltzmann machine and empirical Bayes method

2.1. Boltzmann machine and prior distributions

Consider a fully-connected Boltzmann machine with n Ising variables $\boldsymbol{S} := \{S_i \in \{-1,+1\} \mid i = 1,2,\ldots,n\}$ [1]:

$$P(\boldsymbol{S} \mid h, \boldsymbol{J}) := \frac{1}{Z(h,\boldsymbol{J})} \exp\left(h\sum_{i=1}^{n} S_i + \sum_{i<j} J_{ij}S_iS_j\right), \tag{1}$$

where $\sum_{i<j}$ is the sum over all the distinct pairs of variables; i.e. $\sum_{i<j} = \sum_{i=1}^{n}\sum_{j=i+1}^{n}$. $Z(h,\boldsymbol{J})$ is the partition function defined by

$$Z(h,\boldsymbol{J}) := \sum_{\boldsymbol{S}} \exp\left(h\sum_{i=1}^{n} S_i + \sum_{i<j} J_{ij}S_iS_j\right),$$

where $\sum_{\boldsymbol{S}}$ is the sum over all the possible configurations of $\boldsymbol{S}$; i.e. $\sum_{\boldsymbol{S}} := \prod_{i=1}^{n}\sum_{S_i=\pm 1}$. The parameters, $h \in (-\infty,+\infty)$ and $\boldsymbol{J} := \{J_{ij} \in (-\infty,+\infty) \mid i<j\}$, denote the external field and couplings, respectively.

Given N observed data points, $\mathcal{D} := \{\mathbf{S}^{(\mu)} \in \{-1,+1\}^n \mid \mu = 1,2,\ldots,N\}$, we define the log-likelihood function:

$$L_{\mathrm{ML}}(h,\boldsymbol{J}) := \frac{1}{nN}\sum_{\mu=1}^{N} \ln P(\mathbf{S}^{(\mu)} \mid h,\boldsymbol{J}). \tag{2}$$

Maximizing the log-likelihood function with respect to h and $\boldsymbol{J}$ (i.e. the ML estimation) just corresponds to the BML (or the inverse Ising problem), i.e.

$$\{\hat{h}_{\mathrm{ML}}, \hat{\boldsymbol{J}}_{\mathrm{ML}}\} = \arg\max_{h,\boldsymbol{J}} L_{\mathrm{ML}}(h,\boldsymbol{J}). \tag{3}$$

Now, we introduce prior distributions for the parameters h and $\boldsymbol{J}$ as $P_{\mathrm{prior}}(h \mid H)$ and

$$P_{\mathrm{prior}}(\boldsymbol{J} \mid \gamma) := \prod_{i<j} P_{\mathrm{prior}}(J_{ij} \mid \gamma), \tag{4}$$

respectively. H and γ are the hyperparameters of these prior distributions. One of the most important motivations for introducing the prior distributions is for a Bayesian interpretation of the regularized ML estimation [15]. Given the observed dataset $\mathcal{D}$, by using the prior distributions, the posterior distribution of h and $\boldsymbol{J}$ is expressed as

$$P_{\mathrm{post}}(h,\boldsymbol{J} \mid \mathcal{D}, H, \gamma) = \frac{P(\mathcal{D} \mid h,\boldsymbol{J})P_{\mathrm{prior}}(h \mid H)P_{\mathrm{prior}}(\boldsymbol{J} \mid \gamma)}{P(\mathcal{D} \mid H, \gamma)}, \tag{5}$$

where

$$P(\mathcal{D} \mid h,\boldsymbol{J}) := \prod_{\mu=1}^{N} P(\mathbf{S}^{(\mu)} \mid h,\boldsymbol{J}).$$

The distribution in the denominator in equation (5), $P(\mathcal{D} \mid H, \gamma)$, is sometimes referred to as the evidence. By using the posterior distribution, the maximum *a posteriori* (MAP) estimation of the parameters is obtained as

$$\{\hat{h}_{\mathrm{MAP}}, \hat{\boldsymbol{J}}_{\mathrm{MAP}}\} = \arg\max_{h,\boldsymbol{J}} L_{\mathrm{MAP}}(h, \boldsymbol{J}), \tag{6}$$

where

$$\begin{aligned} L_{\mathrm{MAP}}(h, \boldsymbol{J}) &:= \frac{1}{nN} \ln P_{\mathrm{post}}(h, \boldsymbol{J} \mid \mathcal{D}, H, \gamma) \\ &= L_{\mathrm{ML}}(h, \boldsymbol{J}) + \frac{1}{nN} R_0(h) + \frac{1}{nN} R_1(\boldsymbol{J}) + \text{constant}. \end{aligned} \tag{7}$$

The MAP estimation in equation (6) corresponds to the regularized ML estimation, in which $R_0(h) := \ln P_{\mathrm{prior}}(h \mid H)$ and $R_1(\boldsymbol{J}) := \ln P_{\mathrm{prior}}(\boldsymbol{J} \mid \gamma)$ work as penalty terms. For example, (i) when the prior distribution of $\boldsymbol{J}$ is the Gaussian prior,

$$P_{\mathrm{prior}}(J_{ij} \mid \gamma) = \sqrt{\frac{n}{2\pi\gamma}} \exp\Big(- \frac{nJ_{ij}^2}{2\gamma} \Big), \quad \gamma > 0, \tag{8}$$

$R_1(\boldsymbol{J})$ corresponds to the L_2 regularization term, and γ corresponds to its coefficient; (ii) when the prior distribution of $\boldsymbol{J}$ is the Laplace prior,

$$P_{\mathrm{prior}}(J_{ij} \mid \gamma) = \sqrt{\frac{n}{2\gamma}} \exp\Big(- \sqrt{\frac{2n}{\gamma}} |J_{ij}| \Big), \quad \gamma > 0 \tag{9}$$

$R_1(\boldsymbol{J})$ corresponds to the L_1 regularization term, and γ again corresponds to its coefficient. The variances of these prior distributions are identical, $\mathrm{Var}[J_{ij}] = \gamma/n$. In this study, as a simple test case, we use these two prior distributions for $\boldsymbol{J}$ and

$$P_{\mathrm{prior}}(h \mid H) = \delta(h - H), \tag{10}$$

where $\delta(x)$ is the Dirac delta function, for h.

2.2. Framework of the empirical Bayes method

Using the empirical Bayes method, we can infer the values of the hyperparameters, H and γ, from the observed dataset $\mathcal{D}$. We define a marginal log-likelihood function as

$$L_{\mathrm{EB}}(H, \gamma) := \frac{1}{nN} \ln \big[P(\mathcal{D} \mid h, \boldsymbol{J}) \big]_{h,\boldsymbol{J}}, \tag{11}$$

where $[\cdots]_{h,\boldsymbol{J}}$ is the average over the prior distributions; i.e.

$$[\cdots]_{h,\boldsymbol{J}} := \int \mathrm{d}\boldsymbol{J} \int \mathrm{d}h (\cdots) P_{\mathrm{prior}}(h \mid H) P_{\mathrm{prior}}(\boldsymbol{J} \mid \gamma).$$

We refer to the marginal log-likelihood function as the *empirical Bayes likelihood function* in this study. From the perspective of the empirical Bayes method, the optimal values of the hyperparameters, $\hat{H}$ and $\hat{\gamma}$, are obtained by maximizing of the empirical Bayes likelihood function; i.e.

$$\{\hat{H}, \hat{\gamma}\} = \arg\max_{H,\gamma} L_{\mathrm{EB}}(H, \gamma). \tag{12}$$

It is noteworthy that $[P(\mathcal{D} \mid h, \boldsymbol{J})]_{h,\boldsymbol{J}}$ in equation (11) is identified as the evidence appearing in equation (5).

The marginal log-likelihood function can be rewritten as

$$L_{\mathrm{EB}}(H,\gamma) = \frac{1}{nN} \ln \left[\exp\left(nNL_{\mathrm{ML}}(h,\boldsymbol{J}) \right) \right]_{h,\boldsymbol{J}}. \tag{13}$$

Consider the case $N \gg n$. In this case, by using the saddle point evaluation, equation (13) is reduced to

$$L_{\mathrm{EB}}(H,\gamma) \approx \frac{1}{nN} \ln P_{\mathrm{prior}}(\hat{h}_{\mathrm{ML}} \mid H) + \frac{1}{nN} \ln P_{\mathrm{prior}}(\hat{\boldsymbol{J}}_{\mathrm{ML}} \mid \gamma) + \text{constant}.$$

In this case, the empirical Bayes' estimates $\{\hat{H}, \hat{\gamma}\}$ thus converge to the ML estimates of the hyperparameters in the prior distributions in which the ML estimates of the parameters $\{\hat{h}_{\mathrm{ML}}, \hat{\boldsymbol{J}}_{\mathrm{ML}}\}$ (i.e. the solution to the BML) are inserted. This indicates that the parameter estimations can be conducted independently of the hyperparameter estimation. In this study, we do not concern ourselves with this trivial case.

As discussed, the hyperparameters correspond to the regularization coefficients in the regularized ML estimation. Generally, the appropriate values of the regularization coefficients can be estimated using two methods. First method is the cross validation (CV) method, such as the leave-one-out CV. The CV method is usually used with the regularized ML estimation (or the MAP estimation). In the CV method, the validity of the solution to the regularized ML estimation is checked using the validation dataset separated from the training dataset, and the values of the regularization coefficients providing the optimal solution is selected. This method is effective when the validity of the solution can be checked easily, that is the case in e.g. regression or pattern recognition problems. However, there are some applications in which the validity cannot be checked easily. The graph mining problem [18, 19] is such an example, because we do not know what structure is the best in advance. The second method is the empirical Bayes approach discussed in this paper. This approach allows us to estimate the appropriate hyperparameters in cases where the CV method cannot be used directly, because it is basically based on the fully Bayesian treatment and does not require explicit validation. The empirical Bayes approach is thus important for problems involving difficulties in validation.

3. Statistical mechanical analysis

The empirical Bayes likelihood function in equation (11) involves intractable multiple integrations. In this section, we evaluate the empirical Bayes likelihood function using a statistical mechanical analysis. We consider the two types of the prior distribution of $\boldsymbol{J}$: one is the Gaussian prior in equation (8), and the other is the Laplace prior in equation (9).

First, we evaluate the empirical Bayes likelihood function on the basis of the Gaussian prior in sections 3.1–3.3, after which we describe the evaluation based on the Laplace prior in section 3.4.

3.1. Replica method

The empirical Bayes likelihood function in equation (11) can be represented as

$$L_{\mathrm{EB}}(H,\gamma) = \frac{1}{nN} \ln \lim_{x \to -1} \Psi_x(H,\gamma), \tag{14}$$

where

$$\Psi_x(H,\gamma) := \Big[Z(h,\boldsymbol{J})^{xN} \exp N\Big(h\sum_{i=1}^{n} d_i + \sum_{i<j} J_{ij}d_{ij}\Big)\Big]_{h,\boldsymbol{J}}, \tag{15}$$

and

$$d_i := \frac{1}{N}\sum_{\mu=1}^{N} \mathrm{S}_i^{(\mu)}, \quad d_{ij} := \frac{1}{N}\sum_{\mu=1}^{N} \mathrm{S}_i^{(\mu)}\mathrm{S}_j^{(\mu)}$$

are the sample averages of the observed data points. We assume that $\tau_x := xN$ is a natural number, and therefore equation (15) can be expressed as

$$\begin{aligned}\Psi_x(H,\gamma) = \Big[\sum_{\mathcal{S}_x} \exp\Big\{h\sum_{i=1}^{n}\Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} + Nd_i\Big) \\ + \sum_{i<j} J_{ij}\Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}}S_j^{\{a\}} + Nd_{ij}\Big)\Big\}\Big]_{h,\boldsymbol{J}},\end{aligned} \tag{16}$$

where $a, b \in \{1, 2, \ldots, \tau_x\}$ are replica indices, and $S_i^{\{a\}}$ is the Ising variable on site i in the ath replica. $\mathcal{S}_x := \{S_i^{\{a\}} \mid i = 1, 2, \ldots, n;\ a = 1, 2, \ldots, \tau_x\}$ is the set of all the Ising variables in the replicated system, and $\sum_{\mathcal{S}_x}$ is the sum over all the possible configurations of $\mathcal{S}_x$; i.e. $\sum_{\mathcal{S}_x} := \prod_{i=1}^{n}\prod_{a=1}^{\tau_x}\sum_{S_i^{\{a\}}=\pm 1}$. We evaluate $\Psi_x(H,\gamma)$ under the assumption that τ_x us a natural number, after which we take the limit of $x \to -1$ of the evaluation result to obtain the empirical Bayes likelihood function (this is the so-called *replica trick*).

By employing the Gaussian prior in equation (8), equation (16) becomes

$$\Psi_x^{\mathrm{Gauss}}(H,\gamma) = \exp\Big\{nNHM + \frac{\gamma(n-1)N^2}{4}\Big(C_2 + \frac{x}{N}\Big) - F_x(H,\gamma)\Big\}, \tag{17}$$

where

$$M := \frac{1}{n}\sum_{i=1}^{n} d_i, \quad C_k := \frac{2}{n(n-1)}\sum_{i<j} d_{ij}^k, \tag{18}$$

and

$$F_x(H,\gamma) := -\ln\sum_{\mathcal{S}_x} \exp\big(-E_x(\mathcal{S}_x;H,\gamma)\big) \tag{19}$$

is the replicated (Helmholtz) free energy [20–23]; here,

$$\begin{aligned}E_x(\mathcal{S}_x;H,\gamma) := -H\sum_{i=1}^{n}\sum_{a=1}^{\tau_x} S_i^{\{a\}} - \frac{\gamma N}{n}\sum_{i<j} d_{ij}\sum_{a=1}^{\tau_x} S_i^{\{a\}}S_j^{\{a\}} \\ - \frac{\gamma}{n}\sum_{i<j}\sum_{a<b} S_i^{\{a\}}S_j^{\{a\}}S_i^{\{b\}}S_j^{\{b\}}\end{aligned} \tag{20}$$

is the Hamiltonian of the replicated system, where $\sum_{a<b}$ is the sum over all the distinct pairs of replicas; i.e. $\sum_{a<b} = \sum_{a=1}^{\tau_x}\sum_{b=a+1}^{\tau_x}$.

3.2. Plefka expansion

Because the replicated free energy in equation (19) includes intractable multiple summations, an approximation is needed to proceed with our evaluation. In this section, we approximate the replicated free energy using the Plefka expansion [3]. In brief, the Plefka expansion is the perturbative expansion in a Gibbs free energy that is a dual form of a corresponding Helmholtz free energy.

The Gibbs free energy is obtained as

$$G_x(m,H,\gamma) = -n\tau_x Hm + \operatorname*{extr}_{\lambda}\Big\{\lambda n\tau_x m - \ln\sum_{\mathcal{S}_x}\exp\big(-E_x(\mathcal{S}_x;\lambda,\gamma)\big)\Big\}. \tag{21}$$

The derivation of this Gibbs free energy is described in appendix A. It is noteworthy that this type of expression of the Gibbs free energy implies the replica-symmetric (RS) assumption. To take the replica-symmetry breaking (RSB) into account, explicit treatments of overlaps between different replicas are needed [21]. By expanding $G_x(m,H,\gamma)$ around $\gamma = 0$, we obtain

$$\frac{G_x(m,H,\gamma)}{nN} = -xHm + xe(m) + \phi_x^{(1)}(m)\gamma + \phi_x^{(2)}(m)\gamma^2 + O(\gamma^3), \tag{22}$$

where $e(m)$ is the negative mean-field entropy defined by

$$e(m) := \frac{1+m}{2}\ln\frac{1+m}{2} + \frac{1-m}{2}\ln\frac{1-m}{2}, \tag{23}$$

and the coefficients, $\phi_x^{(1)}(m)$ and $\phi_x^{(2)}(m)$, are expressed as equations (B.4) and (B.9), respectively. The detailed derivation of these coefficients is presented in appendix B.

From equations (14), (17), (22) and (A.4), we obtain the empirical Bayes likelihood function as

$$L_{\mathrm{EB}}(H,\gamma) \approx HM - \operatorname*{extr}_{m}\Big[Hm - e(m) + \Phi(m)\gamma + \phi_{-1}^{(2)}(m)\gamma^2\Big] \tag{24}$$

where

$$\Phi(m) := \phi_{-1}^{(1)}(m) - \frac{(n-1)N}{4n}\Big(C_2 - \frac{1}{N}\Big).$$

From equations (B.4) and (B.9), $\Phi(m)$ and $\phi_{-1}^{(2)}(m)$ are

$$\Phi(m) = \frac{(n-1)NC_1}{2n}m^2 - \frac{(n-1)N}{4n}\Big\{C_2 + \frac{N+1}{N}\Big(m^4 - \frac{1}{N+1}\Big)\Big\} \tag{25}$$

and

$$\begin{aligned}\phi_{-1}^{(2)}(m) = {}& \frac{(n-1)^2N^2\Omega}{2n^2}m^2(1-m^2) + \frac{(n-1)N^2C_2}{4n^2}(1-m^2)^2 \\ & - \frac{(n-1)N(N+1)C_1}{2n^2}m^2(1-m^2)^2 \\ & - \frac{(n-1)(N+1)}{4n^2}(n-N-3)m^4(1-m^2)^2 \\ & - \frac{(n-1)(N+1)}{8n^2}(1-m^4)^2,\end{aligned} \tag{26}$$

respectively. The coefficient Ω appearing in the above equation is defined by

Table 1. Detailed values (averages and standard deviations) of some plots in figure 2 (when $H_{\text{true}} = 0$ and $\alpha = 0.4$).

		J_{true}						
		0	0.2	0.4	0.6	0.8	1	1.2
$\hat{J}$	$n = 300$	0.048 ± 0.06	0.20 ± 0.04	0.41 ± 0.02	0.62 ± 0.02	0.82 ± 0.02	0.96 ± 0.02	1.03 ± 0.02
	$n = 500$	0.038 ± 0.05	0.20 ± 0.03	0.40 ± 0.01	0.62 ± 0.01	0.82 ± 0.01	0.96 ± 0.01	1.03 ± 0.01

$$\Omega := \frac{1}{n}\sum_{i=1}^{n}\omega_i^2, \tag{27}$$

where

$$\omega_i := \frac{1}{n-1}\sum_{j\in\partial(i)} d_{ij} - C_1; \tag{28}$$

here, $\partial(i) := \{1, 2, \ldots, n\} \setminus \{i\}$.

3.3. Inference algorithm

As mentioned in section 2.2, the empirical Bayes inference is achieved by maximizing $L_{\text{EB}}(H, \gamma)$ with respect to H and γ (see equation (12)). From the extremum condition of equation (24) with respect to H, we obtain

$$\hat{m} = M, \tag{29}$$

where $\hat{m}$ is the value of m that satisfies the extremum condition in equation (24). From the extremum condition of equation (24) with respect to m and equation (29), we obtain

$$\hat{H} = \tanh^{-1} M - \left(\frac{\partial\phi_{-1}^{(1)}(m)}{\partial m}\gamma + \frac{\partial\phi_{-1}^{(2)}(m)}{\partial m}\gamma^2\right)\Big|_{m=M}. \tag{30}$$

From equations (24) and (29), the optimal value of γ is obtained by

$$\hat{\gamma} = \arg\max_{\gamma}\left[-\Phi(M)\gamma - \phi_{-1}^{(2)}(M)\gamma^2\right]. \tag{31}$$

From equation (31), $\hat{\gamma}$ is immediately obtained as follows: (i) when $\phi_{-1}^{(2)}(M) > 0$ and $\Phi(M) \geqslant 0$ or when $\phi_{-1}^{(2)}(M) = 0$ and $\Phi(M) > 0$, $\hat{\gamma} = 0$, (ii) when $\phi_{-1}^{(2)}(M) > 0$ and $\Phi(M) < 0$, $\hat{\gamma} = -\Phi(M)/(2\phi_{-1}^{(2)}(M))$, and (iii) $\hat{\gamma} \to \infty$ elsewhere. Here, we ignore the case $\phi_{-1}^{(2)}(M) = \Phi(M) = 0$, because it hardly occurs in realistic settings. By using equations (30) and (31), we can obtain the solution to the empirical Bayes inference without any iterative processes. The pseudocode of the proposed procedure is shown in algorithm 1.

Algorithm 1. Proposed inference algorithm.

1: **Input** Observed data set: $\mathcal{D} := \{\mathbf{S}^{(\mu)} \in \{-1,+1\}^n \mid \mu = 1,2,\ldots,N\}$.
2: Compute M, Ω, C_1, and C_2 using the data set according to equations (18) and (27).
3: Determine $\hat{\gamma}$ using equation (31):

$$\hat{\gamma} = \begin{cases} 0 & \text{case (i)} \\ -\Phi(M)/(2\phi_{-1}^{(2)}(M)) & \text{case (ii)} \\ \infty & \text{elsewhere} \end{cases},$$

where case (i): $\phi_{-1}^{(2)}(M) > 0,\ \Phi(M) \geqslant 0$ or $\phi_{-1}^{(2)}(M) = 0,\ \Phi(M) > 0$ and case (ii):
$\phi_{-1}^{(2)}(M) > 0,\ \Phi(M) < 0$.
4: Using $\hat{\gamma}$, determine $\hat{H}$ using equation (30).
5: **Output** $\hat{\gamma}$ and $\hat{H}$.

In the proposed method, the value of $\hat{H}$ does not affect the determination of $\hat{\gamma}$. Many mean-field-based methods for BML (e.g. listed in section 1) have similar procedures, in which $\hat{\boldsymbol{J}}_{\mathrm{ML}}$ are determined separately from $\hat{h}_{\mathrm{ML}}$. This is seen as one of the common properties of the mean-field-based methods for BML including the current empirical Bayes problem.

3.4. Evaluation based on Laplace prior

The above evaluation was for the Gaussian prior in equation (8). Here, we explain the evaluation for the Laplace prior in equation (9). By employing the Laplace prior in equation (9), equation (16) becomes

$$\Psi_x^{\mathrm{Laplace}}(H,\gamma) = \xi^{n(n-1)} \mathrm{e}^{nNHM} \sum_{\mathcal{S}_x} \exp\Bigg[H \sum_{i=1}^{n} \sum_{a=1}^{\tau_x} S_i^{\{a\}} - \sum_{i<j} \ln\Big\{\xi^2 - \Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} S_j^{\{a\}} + Nd_{ij}\Big)^2\Big\}\Bigg], \tag{32}$$

where $\xi := \sqrt{2n/\gamma}$. Here, we assume

$$\xi > \max_{i<j} \Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} S_j^{\{a\}} + Nd_{ij}\Big). \tag{33}$$

By using the perturbative approximation,

$$\begin{aligned} \ln\Big\{\xi^2 - \Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} S_j^{\{a\}} + Nd_{ij}\Big)^2\Big\} &= \ln \xi^2 + \ln\Big\{1 - \xi^{-2}\Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} S_j^{\{a\}} + Nd_{ij}\Big)^2\Big\} \\ &\approx \ln \xi^2 - \xi^{-2}\Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} S_j^{\{a\}} + Nd_{ij}\Big)^2, \end{aligned}$$

we obtain the approximation of equation (32) as

$$\Psi_x^{\mathrm{Laplace}}(H,\gamma) \approx \mathrm{e}^{nNHM} \sum_{\mathcal{S}_x} \exp\Bigg[H \sum_{i=1}^{n} \sum_{a=1}^{\tau_x} S_i^{\{a\}} + \xi^2 \sum_{i<j} \Big(\sum_{a=1}^{\tau_x} S_i^{\{a\}} S_j^{\{a\}} + Nd_{ij}\Big)^2\Bigg].$$

The right-hand side of this equation coincides with $\Psi_x^{\mathrm{Gauss}}(H,\gamma)$ in equation (17). This means that the empirical Bayes inference based on the Laplace prior in equation (9) is (approximately)

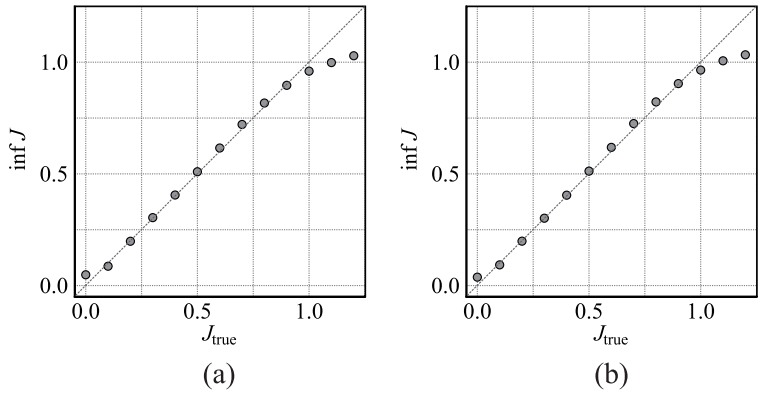

Figure 2. Scatter plots of J_{true} (horizontal axis) versus $\hat{J}$ (vertical axis) when $H_{\text{true}} = 0$ and $\alpha = 0.4$: (a) $n = 300$ and (b) $n = 500$. Plots are the average values over 300 experiments.

equivalent to that based on the Gaussian prior in equation (8) (i.e. $\Psi_x^{\text{Laplace}}(H, \gamma) \approx \Psi_x^{\text{Gauss}}(H, \gamma)$) when the assumption of equation (33) is justified. Thus, we can also use the algorithm presented in section 3.3 for the case of the Laplace prior.

4. Numerical experiments

In this section, we describe the results of our numerical experiments. In these experiments, the observed dataset $\mathcal{D}$ are generated from the generative Boltzmann machine, which has the same form as equation (1), by using annealed importance sampling (AIS) [24]. In AIS, we control the annealing schedule using a series of inverse temperature $0 = \beta_0 < \beta_1 < \cdots < \beta_T < \beta_{\text{final}} = 1$, where we used the annealing schedule of $\beta_{t+1} = \beta_t + 0.03$. The parameters of the generative Boltzmann machine are drawn from the prior distributions in equations (4) and (10). That is, we consider the model-matched case (i.e. the generative and learning models are identical).

In the following, we use the notations $\alpha := N/n$ and $J := \sqrt{\gamma}$. The standard deviations of the Gaussian prior in equation (8) and of the Laplace prior in equation (9) are then $J/\sqrt{n}$. We express the hyperparameters for the generative Boltzmann machine by H_{true} and J_{true}.

4.1. Gaussian prior case

Here, we consider the case in which the prior distribution of $\boldsymbol{J}$ is the Gaussian prior in equation (8). In this case, the Boltzmann machine corresponds to the Sherrington-Kirkpatrick (SK) model [25], and therefore it shows the spin-glass transition at $J = 1$ when $h = 0$ (i.e. when $H = 0$).

First, we consider the case $H_{\text{true}} = 0$. We show the scatter plots for the estimation of $\hat{J}$ for various J_{true} when $H_{\text{true}} = 0$ and $\alpha = 0.4$ in figure 2. The detailed values of the plots for some J_{true} values are shown in table 1.

When $J_{\text{true}} < 1$, our estimates of $\hat{J}$ are in good agreement with J_{true}. This implies that the validity of our perturbative approximation is lost in the spin-glass phase, as is often the case with many mean-field approximations. Figure 3 shows the scatter plots for various α. A smaller α causes $\hat{J}$ to be overestimated and a larger α causes it to be underestimated. At

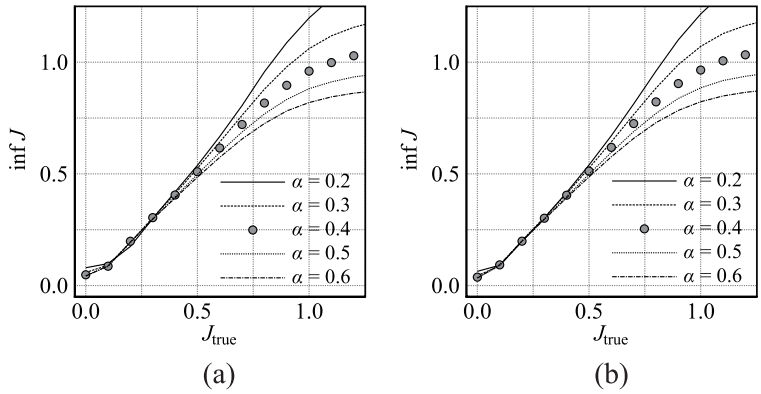

Figure 3. Scatter plots of J_{true} (horizontal axis) versus $\hat{J}$ (vertical axis) for various $\alpha = N/n$ when $H_{\text{true}} = 0$: (a) $n = 300$ and (b) $n = 500$. Plots are the average values over 300 experiments.

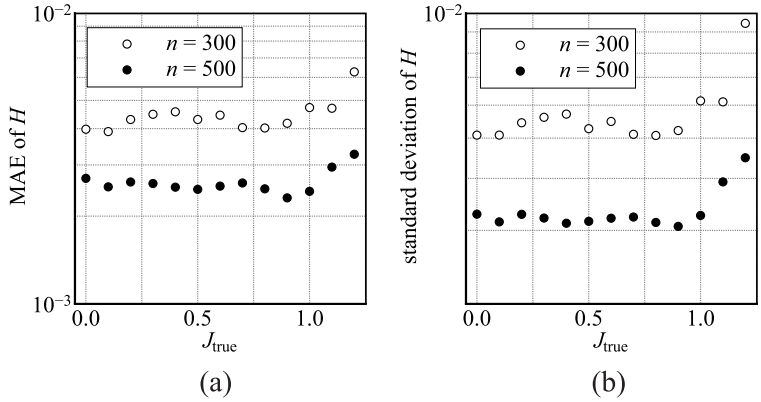

Figure 4. Results of estimation of $\hat{H}$ against J_{true} when $H_{\text{true}} = 0$ and $\alpha = 0.4$: (a) the MAE and (b) standard deviation. Plots are the average values over 300 experiments.

least in our experiments, the optimal value of α seems to be $\alpha_{\text{opt}} \approx 0.4$ when $H_{\text{true}} = 0$. Our method can estimate $\hat{H}$ together with $\hat{J}$. The results for the estimation of $\hat{H}$ when $H_{\text{true}} = 0$ and $\alpha = 0.4$ are shown in figure 4. Figures 4(a) and (b) show the average of $|H_{\text{true}} - \hat{H}|$ (i.e. the mean absolute error (MAE)) and the standard deviation of $\hat{H}$ over 300 experiments, respectively. The MAE and standard deviation increase in the region $J_{\text{true}} > 1$.

Next, we consider the cases $H_{\text{true}} > 0$. The scatter plots for the estimation of $\hat{J}$ for various J_{true} values when $H_{\text{true}} = 0.2$ and $H_{\text{true}} = 0.4$ are shown in figure 5. The appropriate values of α when $H_{\text{true}} = 0.2$ and $H_{\text{true}} = 0.4$ 'approximately' seem to be $\alpha_{\text{opt}} \approx 30/n$ and $\alpha_{\text{opt}} \approx 5/n$, respectively. The detailed values of these plots for some J_{true} values are shown in tables 2 and 3. The results for the estimation of $\hat{H}$ when $H_{\text{true}} = 0.2$ and $\alpha = 30/n$ and when $H_{\text{true}} = 0.4$ and $\alpha = 5/n$ are shown in figures 6 and 7, respectively. The increases in the MAE and standard deviations occur earlier than for the case in figure 4. In the two experiments here, the optimal α is scaled by $O(1/n)$ with respect to n, namely, $N = O(1)$. $N = O(1)$ seems to be too small to estimate the appropriate values of parameters. Regardless,

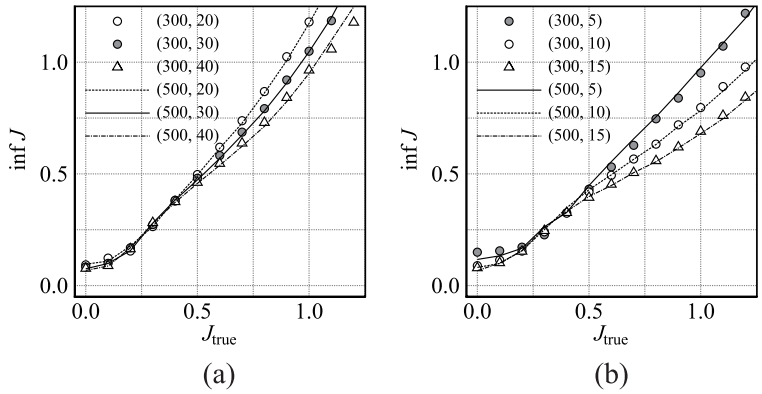

Figure 5. Scatter plots of J_{true} (horizontal axis) versus $\hat{J}$ (vertical axis) for various $\alpha = N/n$ for $n = 300$ and 500: (a) $H_{\text{true}} = 0.2$ and (b) $H_{\text{true}} = 0.4$. Plots are the average values over 300 experiments. The notation in the legend means (n, N).

Table 2. Detailed values (averages and standard deviations) of some plots in figure 5(a) (when $H_{\text{true}} = 0.2$ and $\alpha = 30/n$).

		J_{true}						
		0	0.2	0.4	0.6	0.8	1	1.2
$\hat{J}$	$n = 300$	0.083 ± 0.10	0.17 ± 0.12	0.38 ± 0.07	0.58 ± 0.05	0.79 ± 0.06	1.05 ± 0.12	1.35 ± 0.16
	$n = 500$	0.075 ± 0.09	0.16 ± 0.11	0.38 ± 0.06	0.57 ± 0.04	0.78 ± 0.06	1.05 ± 0.10	1.39 ± 0.16

Table 3. Detailed values (averages and standard deviations) of some plots in figure 5(b) (when $H_{\text{true}} = 0.4$ and $\alpha = 5/n$).

		J_{true}						
		0	0.2	0.4	0.6	0.8	1	1.2
$\hat{J}$	$n = 300$	0.15 ± 0.17	0.17 ± 0.17	0.33 ± 0.19	0.53 ± 0.14	0.75 ± 0.12	0.95 ± 0.14	1.22 ± 0.20
	$n = 500$	0.12 ± 0.15	0.17 ± 0.17	0.33 ± 0.17	0.55 ± 0.12	0.76 ± 0.10	0.98 ± 0.11	1.20 ± 0.16

the results obtained by our method seem to be reasonable. This can be understood as follows. Our method depends on the dataset only through the global statistics of data points, i.e. M, C_k, and Ω. These statistics are regarded as the sample average of the spatial average over the entire space. For example, $M = N^{-1}\sum_{\mu=1}^{N}(n^{-1}\sum_{i=1}^{n} \mathrm{S}_i^{(\mu)})$ can be regarded as the sample average of $M_\mu := n^{-1}\sum_{i=1}^{n} \mathrm{S}_i^{(\mu)}$. M_μ is expected to have the self-averaging property; thus, its variance is negligible for $n \gg 1$. In this case, M takes approximately the same value regardless of the size of N, and therefore, M can have sufficient information to achieve the appropriate estimation even though N is small.

One of the largest qualitative differences between the cases $H_{\text{true}} = 0$ and $H_{\text{true}} > 0$ is the scale of α. In the case $H_{\text{true}} = 0$, the optimal α was scaled by $O(1)$ with respect to n (i.e. $N = O(n)$). Meanwhile, in the case $H_{\text{true}} > 0$, the optimal α is scaled by $O(1/n)$ with respect to n (i.e. $N = O(1)$). This change of scale can be understood from a scale evaluation for the terms in the empirical Bayes likelihood function in equation (24). The detailed reasoning is given in appendix C.

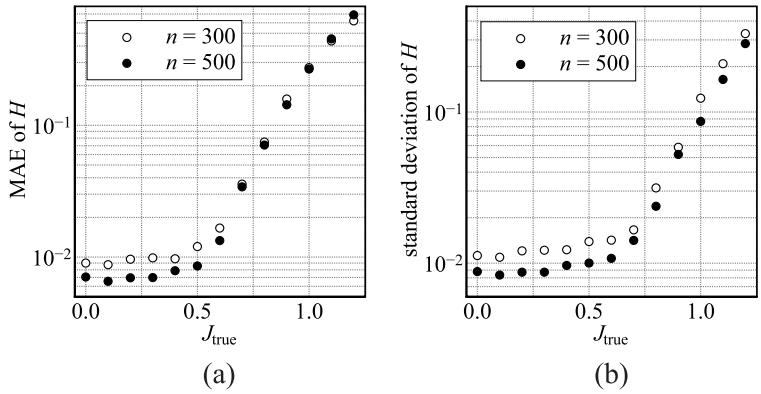

Figure 6. Results of estimation of $\hat{H}$ against J_{true} when $H_{\text{true}} = 0.2$ and $\alpha = 30/n$: (a) the MAE and (b) standard deviation. Plots are the average values over 300 experiments.

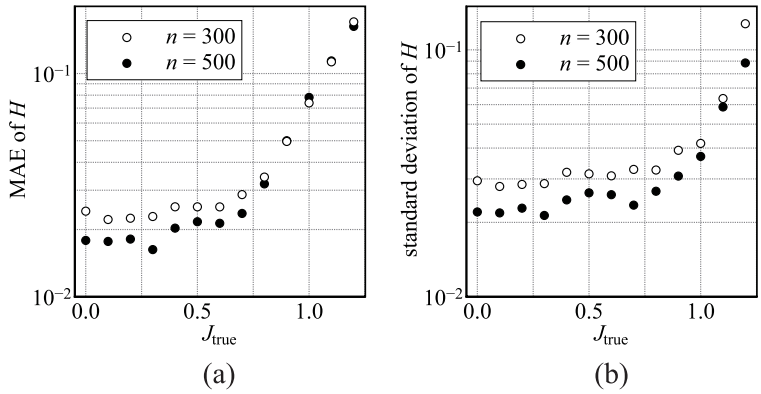

Figure 7. Results of estimation of $\hat{H}$ against J_{true} when $H_{\text{true}} = 0.4$ and $\alpha = 5/n$: (a) the MAE and (b) standard deviation. Plots are the average values over 300 experiments.

4.2. Laplace prior case

Here, we consider the case in which the prior distribution of $\boldsymbol{J}$ is the Laplace prior in equation (9). The scatter plots for the estimation of $\hat{J}$ for various J_{true} values when $H_{\text{true}} = 0$ are shown in figure 8. The plots shown in figure 8 almost completely overlap with those in figure 3. Furthermore, all the numerical results in the case $H_{\text{true}} > 0$ also almost completely overlap with the corresponding results obtained in the above Gaussian prior case, and therefore we do not show those results.

4.3. Comparison with other method

Here, we compare the proposed method with a method based on the maximum pseudo-likelihood estimation (MPLE) [26]. In the MPLE, we approximate the log-likelihood function defined in equation (2) by the pseudo-likelihood function, expressed as

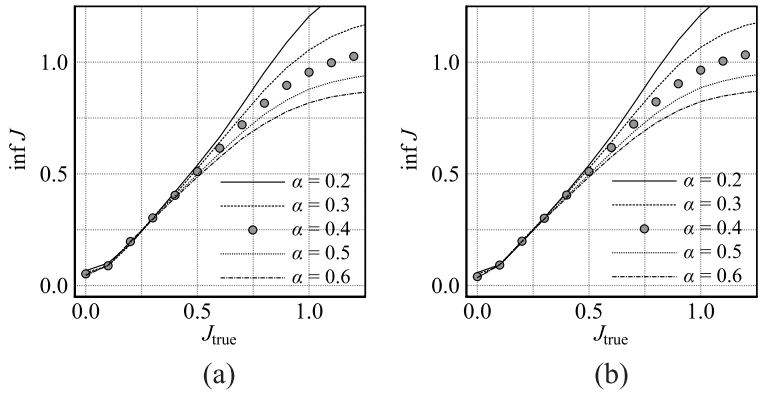

Figure 8. Scatter plots of J_{true} (horizontal axis) versus $\hat{J}$ (vertical axis) for various $\alpha = N/n$, when $H_{\text{true}} = 0$, in the case of the Laplace prior: (a) $n = 300$ and (b) $n = 500$. Plots are the average values over 300 experiments.

$$L_{\text{PL}}(h, \boldsymbol{J}) := hM + \frac{2}{n}\sum_{i<j} J_{ij} d_{ij} - \frac{1}{nN}\sum_{\mu=1}^{N}\sum_{i=1}^{n} \ln 2\cosh\left(h + \sum_{j\in\partial(i)} J_{ij} \mathrm{S}_j^{(\mu)}\right). \quad (34)$$

By using the pseudo-likelihood function, we approximate the empirical Bayes likelihood function in equation (13) as

$$L_{\text{EB}}(H, \gamma) \approx L_{\text{EB}}^{\text{PL}}(H, \gamma) := \frac{1}{nN} \ln \left[\exp\left(nNL_{\text{PL}}(h, \boldsymbol{J})\right) \right]_{h\boldsymbol{J}}. \quad (35)$$

It is known that, when the data points are generated from the same Boltzmann machine, the solution of the MPLE converges to that of the ML estimation for $N \to \infty$ limit [27]. Thus, it can be expected that the solution obtained by the maximization of equation (35) is good when $N \gg 1$.

For simplicity, we ignore the external field h and consider the maximizing problem with only γ: $\hat{\gamma} = \arg\max_\gamma L_{\text{EB}}^{\text{PL}}(\gamma)$, and the prior distribution of $\boldsymbol{J}$ is the Gaussian prior with $J_{\text{true}} = 0.6$. In the experiment, we maximize $L_{\text{EB}}^{\text{PL}}(\gamma)$ numerically based on the line search and evaluate the multiple integrations with respect to $\boldsymbol{J}$ using the Monte Carlo integration with 10^4 samples. The results are summarized in table 4. Unlike our expectation, the estimated $\hat{J}$ becomes worse as N increases. This is presumably because of the Monte Carlo integration used to evaluate the multiple integrations over $\boldsymbol{J}$. The distribution of the integrated function, $\exp(nNL_{\text{PL}}(\boldsymbol{J}))$, is strongly localized around the maximum point of $L_{\text{PL}}(\boldsymbol{J})$ when nN is large. A precise evaluation for such strongly localized distribution by the simple Monte Carlo integration is generally difficult.

From the above observations, we believe that the pseudo-likelihood approach cannot be effective unless a special treatment is introduced for evaluating the multiple integrations.

5. Summary and discussions

In this study, we proposed a hyperparameters inference algorithm by analyzing the empirical Bayes likelihood function in equation (11) using the replica method and the Plefka expansion. The validity of our method was examined in numerical experiments for the Gaussian and

Table 4. Results obtained from maximizing $L_{\mathrm{EB}}^{\mathrm{PL}}(\gamma)$ (averages and standard deviations) for various N when $n = 100$ and $J_{\mathrm{true}} = 0.6$. The values in the table are estimated over 50 experiments. The result obtained from our method is $\hat{J} = 0.61 \pm 0.04$ with $\alpha = 0.4$, which is estimated over 300 experiments.

	N					
	5	10	20	40	80	160
$\hat{J}$	0.25 ± 0.02	0.18 ± 0.02	0.13 ± 0.02	0.09 ± 0.01	0.07 ± 0.01	0.06 ± 0.005

Laplace priors, which demonstrated the existence of an appropriate scale in the size of the dataset that can accurately recover the values of the hyperparameters.

However, some problems remain. The first one is the scale of N. In our experiments, we found that an appropriate N is scaled by $O(n)$ when $H_{\mathrm{true}} = 0$ or by $O(1)$ when $H_{\mathrm{true}} \neq 0$. However, such scales seem to be unnatural, because they should not appear in the original framework of the empirical Bayes method. As discussed in section 2.2, when $N \gg n$, maximizing the empirical Bayes likelihood function is reduced to the ML estimation of the prior distributions for the solution to BML. This must lead to the correct $\hat{\gamma}$ and $\hat{H}$, because the solution to BML is perfect when $N \to \infty$. For reference, we show the results obtained from the original framework. When n is small, we can numerically evaluate $L_{\mathrm{ML}}(h, \boldsymbol{J})$; therefore, we can evaluate equation (13) numerically. Figure 9 shows the MAE between $\hat{J}$ and J_{true} (i.e. the average of $|J_{\mathrm{true}} - \hat{J}|$) for various N when $n = 5$. In the experiment, we ignore h and evaluate the multiple integrations over $\boldsymbol{J}$ using the Monte Carlo integration (with 10^5 samples), as we did in section 4.3. The MAEs monotonically decrease with an increase in N and the unnatural scales, that our method has, does not seem to appear. Therefore, such unnatural scales appear due to our approximation, which is also supported by a scale analysis given in appendix C. An improvement of the approximation (e.g. by evaluating the leading terms in the Plefka expansion or using some other approximations) might reduce these unnatural behaviors.

The second problem is the optimal value of $\alpha = N/n$. Empirically, we found that $\alpha_{\mathrm{opt}} \approx 0.4$ when $H_{\mathrm{true}} = 0$ and that it decreases as H_{true} increases (e.g. $\alpha_{\mathrm{opt}} \approx 30/n$ when $H_{\mathrm{true}} = 0.2$ and $\alpha_{\mathrm{opt}} \approx 5/n$ when $H_{\mathrm{true}} = 0.4$). As can be seen in the results of our experiments, the solution to our method is robust for the choice of α when J_{true} is small ($J_{\mathrm{true}} < J_c$) and is sensitive to it when J_{true} is large ($J_{\mathrm{true}} > J_c$), where $J_c \approx 0.4$. The estimation of α_{opt} is very important for our method, and it will make our method more practical. This problem would be strongly related to the first problem.

The third problem is the degradation of the estimation accuracy in the spin-glass phase. In our experiments, the estimation accuracies of $\hat{\gamma}$ and $\hat{H}$ were obviously degraded in the spin-glass phase. This means that our Plefka expansion based on the RS assumption loses its validity in the spin-glass phase. In [21], a Plefka expansion for the one-step RSB was proposed. Employing this expansion instead of the current expansion could reduce the degradation in the spin-glass phase. These three problems should be addressed in our future studies.

In this study, we used fully-connected Boltzmann machines whose variables are all visible. We are also interested in an extension of our method to other types of Boltzmann machines such as Boltzmann machines having specific structures or hidden variables. Furthermore, we considered the model-matched case (i.e. the case in which the generative mode and learning model are the same model) in the current study, but model-mismatched cases are more practical and important.

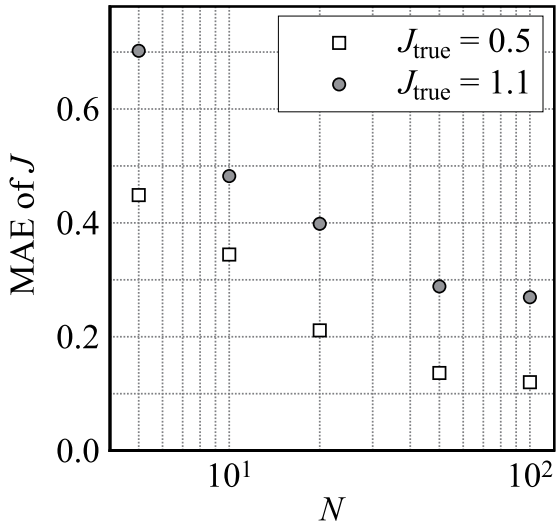

Figure 9. Plots of MAE of $\hat{J}$ obtained from the numerical maximization of equation (13), for $N = 5, 10, 20, 50, 100$. Plots are the average values over 100 experiments.

Acknowledgments

This work was partially supported by JSPS KAKENHI (Grant Nos.: 15H03699, 18K11459, 18H03303, 25120013, 18K11463 and 17H00764), JST CREST (Grant No.: JPMJCR1402), and the COI Program from the JST (Grant No. JPMJCE1312). TO is also supported by a Grant for Basic Science Research Projects from the Sumitomo Foundation.

Appendix A. Gibbs free energy

In this appendix, we derive the Gibbs free energy for the replicated (Helmholtz) free energy in equation (19).

The replicated free energy is obtained by minimizing the variational free energy, defined by

$$f[Q] := \sum_{\mathcal{S}_x} E_x(\mathcal{S}; H, \gamma) Q(\mathcal{S}_x) + \sum_{\mathcal{S}_x} Q(\mathcal{S}_x) \ln Q(\mathcal{S}_x), \tag{A.1}$$

under the normalization constraint, i.e. $\sum_{\mathcal{S}_x} Q(\mathcal{S}_x) = 1$, where $Q(\mathcal{S}_x)$ is a test distribution over $\mathcal{S}_x$, and $E_x(\mathcal{S}_x; H, \gamma)$ is the Hamiltonian for the replicated system defined in equation (20).

The Gibbs free energy is obtained by adding new constraints to the minimization of $f[Q]$. Here, we add the relation $m = (n\tau_x)^{-1} \sum_{i=1}^{n} \sum_{a=1}^{\tau_x} \sum_{\mathcal{S}_x} S_i^{\{a\}} Q(\mathcal{S}_x)$ as the constraint. By using Lagrange multipliers, the Gibbs free energy is obtained as

$$\begin{aligned} G_x(m, H, \gamma) := \operatorname*{extr}_{Q,\lambda,r} \Big\{ & f[Q] - r\Big(\sum_{\mathcal{S}_x} Q(\mathcal{S}_x) - 1 \Big) \\ & - \lambda \Big(\sum_{i=1}^{n} \sum_{a=1}^{\tau_x} \sum_{\mathcal{S}_x} S_i^{\{a\}} Q(\mathcal{S}_x) - n\tau_x m \Big) \Big\}, \end{aligned} \tag{A.2}$$

where 'extr' denotes the extremum with respect to the assigned parameters. By performing the extremum operation with respect to $Q(\mathcal{S})$ and r in equation (A.2), we obtain

$$G_x(m,H,\gamma) = \underset{\lambda}{\mathrm{extr}}\Big\{\lambda n\tau_x m - \ln\sum_{\mathcal{S}_x}\exp\big(-E_x(\mathcal{S}_x;H+\lambda,\gamma)\big)\Big\}. \tag{A.3}$$

The replicated free energy in equation (19) coincides with the extremum of this Gibbs free energy with respect to m; i.e.

$$F_x(H,\gamma) = \underset{m}{\mathrm{extr}}G_x(m,H,\gamma). \tag{A.4}$$

By performing the shift $H+\lambda\to\lambda$ in equation (A.3), we obtain equation (21).

Appendix B. Derivation of coefficients of Plefka expansion

The Plefka expansion considered in this study can be obtained by expanding the Gibbs free energy in equation (21) around $\gamma = 0$.

When $\gamma = 0$, we have

$$\begin{aligned} G_x(m,H,0) &= -n\tau_x Hm + n\tau_x \underset{\lambda}{\mathrm{extr}}\big(\lambda m - \ln 2\cosh\lambda\big) \\ &= -n\tau_x Hm + n\tau_x e(m), \end{aligned} \tag{B.1}$$

where $e(m)$ is defined in equation (23).

For the derivations of the coefficients $\phi_x^{(1)}(m)$ and $\phi_x^{(2)}(m)$, we decompose $E_x(\mathcal{S}_x;H,\lambda)$ in equation (21) into two parts:

$$E_x(\mathcal{S}_x;\lambda,\gamma) = -\lambda\sum_{i=1}^{n}\sum_{a=1}^{\tau_x}S_i^{\{a\}} + \gamma E_x^{\mathrm{int}}(\mathcal{S}_x),$$

where

$$E_x^{\mathrm{int}}(\mathcal{S}_x) := -\frac{N}{n}\sum_{i<j}d_{ij}\sum_{a=1}^{\tau_x}S_i^{\{a\}}S_j^{\{a\}} - \frac{1}{n}\sum_{i<j}\sum_{a<b}S_i^{\{a\}}S_j^{\{a\}}S_i^{\{b\}}S_j^{\{b\}}.$$

Coefficient $\phi_x^{(1)}(m)$ is defined by

$$\phi_x^{(1)}(m) := \frac{1}{nN}\frac{\partial G_x(m,H,\gamma)}{\partial\gamma}\Big|_{\gamma=0}.$$

The derivative leads to

$$\frac{\partial G_x(m,H,\gamma)}{\partial\gamma} = \Big\langle E_x^{\mathrm{int}}(\mathcal{S}_x)\Big\rangle_\gamma, \tag{B.2}$$

where $\langle\cdots\rangle_\gamma$ denotes the average for the distribution

$$P(\mathcal{S}_x\mid\gamma,m)\propto\exp\big(-E_x(\mathcal{S}_x;\lambda^*,\gamma)\big),$$

where λ^* is the value of λ that satisfies the extremum condition in equation (21) and which is the function relating γ and m; i.e. $\lambda^* = \lambda^*(\gamma,m)$. From the extremum condition for λ in equation (21), we obtain the equation

$$m = \frac{1}{n\tau_x}\sum_{i=1}^{n}\sum_{a=1}^{\tau_x}\langle S_i^{\{a\}}\rangle_\gamma, \tag{B.3}$$

which holds for any γ. In the derivation of equation (B.2), we used equation (B.3). When $\gamma = 0$, equation (B.3) reduces to $m = \tanh \lambda^*$. This means that $\langle S_i^{\{a\}} \rangle_0 = m$ for any i and a. Therefore, we obtain

$$\phi_x^{(1)}(m) = -\frac{x(n-1)NC_1}{2n} m^2 - \frac{(n-1)K_x}{2nN} m^4, \tag{B.4}$$

where $K_x := \tau_x(\tau_x - 1)/2$. In the derivation of equation (B.4), we used the relation $\langle S_i^{\{a\}} S_j^{\{b\}} \rangle_0 = \langle S_i^{\{a\}} \rangle_0 \langle S_j^{\{b\}} \rangle_0$ if $i \neq j$ or $a \neq b$.

The coefficient $\phi_x^{(2)}(m)$ is defined by

$$\phi_x^{(2)}(m) := \frac{1}{2nN} \frac{\partial^2 G_x(m, H, \gamma)}{\partial \gamma^2}\Big|_{\gamma=0}.$$

From equation (B.2), the second derivative is

$$\frac{\partial^2 G_x(m, H, \gamma; \mathcal{D})}{\partial \gamma^2} = \frac{\partial}{\partial \gamma} \left\langle E_x^{\text{int}}(\mathcal{S}_x) \right\rangle_\gamma = \left\langle E_x^{\text{int}}(\mathcal{S}_x) U_x(\gamma) \right\rangle_\gamma, \tag{B.5}$$

where

$$U_x(\gamma) := \left\langle \frac{\partial E_x(\mathcal{S}_x; \lambda^*, \gamma)}{\partial \gamma} \right\rangle_\gamma - \frac{\partial E_x(\mathcal{S}_x; \lambda^*, \gamma)}{\partial \gamma}$$

is Georges's operator, proposed in [28]. To simplify the notation, we omit the explicit description of the dependency of the operator on $\mathcal{S}_x$ and m. By using this operator, the derivative of $\langle A \rangle_\gamma$ with respect to γ is obtained as

$$\frac{\partial \langle A \rangle_\gamma}{\partial \gamma} = \left\langle \frac{\partial A}{\partial \gamma} \right\rangle_\gamma + \langle A U_x(\gamma) \rangle_\gamma .$$

This immediately leads to $\langle S_i^{\{a\}} U_x(\gamma) \rangle_\gamma = 0$, because $\partial \langle S_i^{\{a\}} \rangle_\gamma / \partial \gamma = \partial m / \partial \gamma = 0$. Therefore,

$$\left\langle U_x(\gamma)^2 \right\rangle_\gamma = -\left\langle U_x(\gamma) \frac{\partial E_x(\mathcal{S}_x, \lambda^*, \gamma)}{\partial \gamma} \right\rangle_\gamma = -\left\langle E_x^{\text{int}}(\mathcal{S}_x) U_x(\gamma) \right\rangle_\gamma \tag{B.6}$$

is obtained, where we have used $\langle U_x(\gamma) \rangle_\gamma = 0$. From equations (B.5) and (B.6), we have

$$\frac{\partial^2 G_x(m, H, \gamma)}{\partial \gamma^2} = -\left\langle U_x(\gamma)^2 \right\rangle_\gamma . \tag{B.7}$$

Because

$$\frac{\partial \lambda^*}{\partial \gamma}\Big|_{\gamma=0} = \frac{1}{n\tau_x} \frac{\partial}{\partial \gamma} \frac{\partial G_x(m, H, \gamma)}{\partial m}\Big|_{\gamma=0} = \frac{N}{\tau_x} \frac{\partial \phi_x^{(1)}(m)}{\partial m},$$

when $\gamma = 0$, we obtain

$$\begin{aligned} U_x(0) = {} & \frac{(n-1)N}{n} \sum_{i=1}^{n} \omega_i m \sum_{a=1}^{\tau_x} \left(S_i^{\{a\}} - m\right) \\ & + \frac{N}{n} \sum_{i<j} \left(d_{ij} + \frac{\tau_x - 1}{N} m^2 \right) \sum_{a=1}^{\tau_x} \left(S_i^{\{a\}} - m\right)\left(S_j^{\{a\}} - m\right) \\ & + \frac{1}{n} \sum_{i<j} \sum_{a<b} \left(S_i^{\{a\}} S_j^{\{a\}} - m^2\right)\left(S_i^{\{b\}} S_j^{\{b\}} - m^2\right), \end{aligned} \tag{B.8}$$

where ω_i is defined in equation (28).

By using equations (B.7) and (B.8), we obtain

$$\begin{aligned}\phi_x^{(2)}(m) = &-\frac{(n-1)^2\tau_x N\Omega}{2n^2}m^2(1-m^2) - \frac{(n-1)\tau_x NC_2}{4n^2}(1-m^2)^2 \\ &- \frac{(n-1)K_xC_1}{n^2}m^2(1-m^2)^2 - \frac{(n-1)K_x}{2n^2N}\big(n+\tau_x-3\big)m^4(1-m^2)^2 \\ &- \frac{(n-1)K_x}{4n^2N}(1-m^4)^2,\end{aligned} \tag{B.9}$$

where Ω is defined in equation (27).

Appendix C. Evaluation of orders of each term in the empirical Bayes likelihood

Here, we evaluate the orders of each term in equation (24) with $m = M$, with respect to $n \gg 1$, that is, the orders of each term in

$$L_{\mathrm{EB}}(H,\gamma) \approx e(M) - \Phi(M)\gamma - \phi_{-1}^{(2)}(M)\gamma^2. \tag{C.1}$$

In the following, we assume that $N = O\big(n^\rho\big)$ $(\rho \geqslant 0)$ and that $\{\mathrm{S}_i^{(\mu)}\}$ are i.i.d. samples from a certain distribution.

First, we consider the case $H_{\mathrm{true}} = 0$ in which the distribution of $\{\mathrm{S}_i^{(\mu)}\}$ is unbiased. In this case, we obtain $M = O\big(n^{-(1+\rho)/2}\big)$, $C_1 = O\big(n^{-1-\rho/2}\big)$, and

$$C_2 = \frac{1}{N} + \frac{1}{n(n-1)N^2}\sum_{\mu<\nu}\sum_{i<j}\mathrm{S}_i^{(\mu)}\mathrm{S}_j^{(\mu)}\mathrm{S}_i^{(\nu)}\mathrm{S}_j^{(\nu)} = O\big(n^{-\rho}\big).$$

Similarly, we obtain

$$\begin{aligned}\Omega &= \frac{1}{n(n-1)^2N^2}\sum_{i=1}^{n}\sum_{\mu,\nu=1}^{N}\sum_{j,k\in\partial(i)}\mathrm{S}_i^{(\mu)}\mathrm{S}_j^{(\mu)}\mathrm{S}_i^{(\nu)}\mathrm{S}_k^{(\nu)} - C_1^2 \\ &= \frac{1}{(n-1)N} + \frac{1}{n(n-1)^2N^2}\sum_{i=1}^{n}\sum_{\mu=1}^{N}\sum_{j\neq k\in\partial(i)}\mathrm{S}_j^{(\mu)}\mathrm{S}_k^{(\mu)} \\ &\quad + \frac{1}{n(n-1)^2N^2}\sum_{i=1}^{n}\sum_{\mu\neq\nu}\sum_{j,k\in\partial(i)}\mathrm{S}_i^{(\mu)}\mathrm{S}_j^{(\mu)}\mathrm{S}_i^{(\nu)}\mathrm{S}_k^{(\nu)} - C_1^2 \\ &= O\big(n^{-1-\rho}\big),\end{aligned}$$

because $C_1^2 = O\big(n^{-2-\rho}\big)$. Using the above results and equations (23), (25) and (26), we obtain $e(M) = O(1)$, $\Phi(M) = O(1)$, and $\phi_{-1}^{(2)}(M) = O\big(n^{\rho-1}\big)$, respectively. Therefore, when $\rho = 1$, the orders of all the terms in equation (C.1) are just $O(1)$ with respect to n.

Next, we consider the case $H_{\mathrm{true}} \neq 0$ in which the distribution of $\{\mathrm{S}_i^{(\mu)}\}$ is biased. In this case, M, C_1, and C_2 are $O(1)$, and furthermore, Ω is $O(1)$ because $\omega_i = O(1)$. This leads to $e(M) = O(1)$, $\Phi(M) = O\big(n^\rho\big)$, and $\phi_{-1}^{(2)}(M) = O\big(n^{2\rho}\big)$. Therefore, when $\rho = 0$, the orders of all the terms in equation (C.1) are just $O(1)$ with respect to n.

This consideration and the experiments in section 4 imply that our method based on the Plefka expansion can be validated when all the terms in the empirical Bayes likelihood are

$O(1)$. The introduction of the external field changes the condition to satisfy this criterion, leading to the appropriate scaling of α. This statement is consistent with the numerical observation that a stable result is obtained even for different n's as long as the appropriate scale in α is maintained, as shown in section 4.

ORCID iDs

Muneki Yasuda https://orcid.org/0000-0001-5531-9842
Tomoyuki Obuchi https://orcid.org/0000-0003-1216-489X

References

[1] Ackley D H, Hinton G E and Sejnowski T J 1985 *Cogn. Sci.* **9** 147–69
[2] Roudi Y, Aurell E and Hertz J 2009 *Frontiers Computat. Neurosci.* **3** 1–22
[3] Plefka T 1982 *J. Phys. A: Math. Gen.* **15** 1971–8
[4] Pelizzola A 2005 *J. Phys. A: Math. Gen.* **38** R309
[5] Kappen H J and Rodríguez F B 1998 *Neural Comput.* **10** 1137–56
[6] Tanaka T 1998 *Phys. Rev.* E **58** 2302–10
[7] Yasuda M and Horiguchi T 2006 *Physica* A **368** 83–95
[8] Sessak V and Monasson R 2009 *J. Phys. A: Math. Theor.* **42** 055001
[9] Yasuda M and Tanaka K 2009 *Neural Comput.* **21** 3130–78
[10] Ricci-Tersenghi F 2012 *J. Stat. Mech.* P08015
[11] Furtlehner C 2013 *J. Stat. Mech.* P09020
[12] Sohl-Dickstein J, Battaglino P B and DeWeese M R 2011 *Phys. Rev. Lett.* **107** 220601
[13] Yasuda M 2015 *J. Phys. Soc. Japan* **84** 034001
[14] MacKay D J C 1992 *Neural Comput.* **4** 415–47
[15] Bishop C M 2006 *Pattern Recognition and Machine Learning* (Berlin: Springer)
[16] Mezard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond: an Introduction to the Replica Method and Its Applications* (Singapore: World Scientific)
[17] Nishimori H 2001 *Statistical Physics of Spin Glass and Information Processing—Introduction* (Oxford: Oxford University Press)
[18] Friedman J, Hastie T and Tibshirani R 2008 *Biostatistics* **9** 432–41
[19] Yasuda M, Katou K, Mikuni Y, Yokoyama Y, Harada T, Tanaka A and Yokoyama M 2019 *Nonlinear Theory Appl.* **10** 485–95
[20] Rizzo T, Lage-Castellanos A, Mulet R and Ricci-Tersenghi F 2010 *J. Stat. Phys.* **139** 375–416
[21] Yasuda M, Kabashima Y and Tanaka K 2012 *J. Stat. Mech.* P04002
[22] Lage-Castellanos A, Mulet R, Ricci-Tersenghi F and Rizzo T 2013 *J. Phys. A: Math. Theor.* **46** 135001
[23] Yasuda M, Kataoka S and Tanaka K 2015 *Phys. Rev.* E **92** 042120
[24] Neal R M 2001 *Stat. Comput.* **11** 125–39
[25] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792–6
[26] Besag J 1975 *J. R. Stat. Soc.* D **24** 179–95
[27] Hyvärinen A 2006 *Neural Comput.* **18** 2283–92
[28] Georges A and Yedidia J S 1991 *J. Phys. A: Math. Gen.* **24** 2173–92