

Chapter 11

Empirical Bayes Method for Boltzmann Machines



Muneki Yasuda

Abstract The framework of the empirical Bayes method allows the estimation of the values of the hyperparameters in the Boltzmann machine by maximizing a specific likelihood function referred to as the empirical Bayes likelihood function. However, the maximization is computationally difficult because the empirical Bayes likelihood function involves intractable integrations of the partition function. The method presented in this chapter avoids this computational problem by using the replica method and the Plefka expansion, which is quite simple and fast because it does not require any iterative procedures and gives reasonable estimates under certain conditions.

11.1 Introduction

Boltzmann machine learning (BML) [1] has been actively studied in the fields of machine learning and 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 as an Ising model, and it can be treated as an inverse problem for the Ising model. The framework of the *usual* BML is as follows. Given a set of observed data points, the appropriate values of the Boltzmann machine parameters, namely the biases and couplings, are estimated through maximum likelihood (ML) estimation. Because BML involves intractable multiple summations (i.e., evaluation of the partition function), several approximations have been proposed for it from the viewpoint of statistical mechanics [2]. Examples include methods based on mean-field approximations (e.g., the Plefka expansion [3] and the cluster variation method [4]) [5–11] and methods based on other approximations [12–14].

This chapter focuses on another type of learning problem for the Boltzmann machine. Consider the prior distributions of the Boltzmann machine parameters and assume that the prior distributions are governed by some hyperparameters. The introduction of the prior distributions is strongly connected to regularized ML estimation, in which the hyperparameters can be regarded as regularization coefficients. The reg-

M. Yasuda (✉)

Graduate School of Science and Engineering, Yamagata University, Yamagata, Japan
e-mail: muneki@yz.yamagata-u.ac.jp

© The Author(s) 2022
N. Katoh et al. (eds.), *Sublinear Computation Paradigm*,
https://doi.org/10.1007/978-981-16-4095-7_11

277

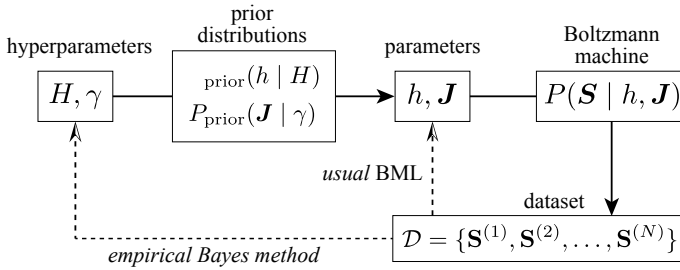


Fig. 11.1 Illustration of scheme of the empirical Bayes method considered in this chapter

ularized ML estimation is important for preventing overfitting to the dataset. As mentioned above, the *usual BML* aims to optimize the values of the Boltzmann machine parameters using a set of observed data points. However, the aim of the problem presented in this chapter is the estimation of the appropriate values of the hyperparameters from the dataset *without* estimating the specific values of the Boltzmann machine parameters. From the Bayesian viewpoint, this can be potentially accomplished by the *empirical Bayes method* (also known as type-II ML estimation or evidence approximation) [15, 16]. The schemes of the *usual BML* and the problem investigated in this chapter are illustrated in Fig. 11.1.

Recently, an effective algorithm was proposed for the empirical Bayes method for the Boltzmann machine [17]. Using this method, the hyperparameter estimates can be obtained without costly operations. This chapter aims to explain this effective method.

The rest of this chapter is organized as follows. The formulations of the Boltzmann machine and its usual and regularized ML estimations are presented in Sect. 11.2. The empirical Bayes method for the Boltzmann machine is presented in Sect. 11.3. Section 11.4 describes a statistical mechanical analysis for the empirical Bayes method and an inference algorithm obtained from the analysis. Experimental results for the presented algorithm are presented in Sect. 11.5. The summary and some discussions are presented in Sect. 11.6. The appendices for this chapter are given in Sect. 11.7.

11.2 Boltzmann Machine with Prior Distributions

Consider a fully connected Boltzmann machine with n (bipole) variables $\mathbf{S} := \{S_i \in \{-1, +1\} \mid i = 1, 2, \dots, n\}$ [1]:

$$P(\mathbf{S} | h, \mathbf{J}) := \frac{1}{Z(h, \mathbf{J})} \exp\left(h \sum_{i=1}^n S_i + \sum_{i < j} J_{ij} S_i S_j\right), \quad (11.1)$$

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

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

where $\sum_{\mathbf{S}}$ is the sum over all possible configurations of \mathbf{S} , that is,

$$\sum_{\mathbf{S}} := \prod_{i=1}^n \sum_{S_i = \pm 1}.$$

The parameters $h \in (-\infty, +\infty)$ and $\mathbf{J} := \{J_{ij} \in (-\infty, +\infty) \mid i < j\}$ denote the bias and couplings, respectively.

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

$$L_{\text{ML}}(h, \mathbf{J}) := \frac{1}{nN} \sum_{\mu=1}^N \ln P(\mathbf{S}^{(\mu)} \mid h, \mathbf{J}). \quad (11.2)$$

The maximization of the log-likelihood function with respect to h and \mathbf{J} (i.e., the ML estimation) corresponds to BML (or the inverse Ising problem), that is,

$$\{\hat{h}_{\text{ML}}, \hat{\mathbf{J}}_{\text{ML}}\} = \arg \max_{h, \mathbf{J}} L_{\text{ML}}(h, \mathbf{J}). \quad (11.3)$$

However, the exact ML estimations cannot be obtained because the gradients of the log-likelihood function include intractable sums over $O(2^n)$ terms.

We now introduce the prior distributions of the parameters h and \mathbf{J} as $P_{\text{prior}}(h \mid H)$ and

$$P_{\text{prior}}(\mathbf{J} \mid \gamma) := \prod_{i<j} P_{\text{prior}}(J_{ij} \mid \gamma), \quad (11.4)$$

where H and γ are the hyperparameters of these prior distributions. One of the most important motivations for introducing the prior distributions is the Bayesian interpretation of the regularized ML estimation [16]. Given the observed dataset \mathcal{D} , using the prior distributions, the posterior distribution of h and \mathbf{J} is expressed as

$$P_{\text{post}}(h, \mathbf{J} \mid \mathcal{D}, H, \gamma) = \frac{P(\mathcal{D} \mid h, \mathbf{J}) P_{\text{prior}}(h \mid H) P_{\text{prior}}(\mathbf{J} \mid \gamma)}{P(\mathcal{D} \mid H, \gamma)}, \quad (11.5)$$

where

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

The denominator of Eq. (11.5) is sometimes referred to as *evidence*. Using the posterior distribution, the maximum a posteriori (MAP) estimation of the parameters is obtained as

$$\{\hat{h}_{\text{MAP}}, \hat{\mathbf{J}}_{\text{MAP}}\} = \arg \max_{h, \mathbf{J}} L_{\text{MAP}}(h, \mathbf{J}), \quad (11.6)$$

where

$$\begin{aligned} L_{\text{MAP}}(h, \mathbf{J}) &:= \frac{1}{nN} \ln P_{\text{post}}(h, \mathbf{J} | \mathcal{D}, H, \gamma) \\ &= L_{\text{ML}}(h, \mathbf{J}) + \frac{1}{nN} R_0(h) + \frac{1}{nN} R_1(\mathbf{J}) + \text{constant}. \end{aligned} \quad (11.7)$$

The MAP estimation of Eq. (11.6) corresponds to the regularized ML estimation, in which $R_0(h) := \ln P_{\text{prior}}(h | H)$ and $R_1(\mathbf{J}) := \ln P_{\text{prior}}(\mathbf{J} | \gamma)$ work as penalty terms. For example, (i) when the prior distribution of \mathbf{J} is a Gaussian prior,

$$P_{\text{prior}}(J_{ij} | \gamma) = \sqrt{\frac{n}{2\pi\gamma}} \exp\left(-\frac{nJ_{ij}^2}{2\gamma}\right), \quad \gamma > 0, \quad (11.8)$$

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

$$P_{\text{prior}}(J_{ij} | \gamma) = \sqrt{\frac{n}{2\gamma}} \exp\left(-\sqrt{\frac{2n}{\gamma}} |J_{ij}|\right), \quad \gamma > 0, \quad (11.9)$$

$R_1(\mathbf{J})$ corresponds to the L_1 regularization term and γ again corresponds to its coefficient. The variances of these prior distributions are identical, that is, $\text{Var}[J_{ij}] = \gamma/n$.

The following uses the Gaussian prior for \mathbf{J} and the following as a simple test case:

$$P_{\text{prior}}(h | H) = \delta(h - H), \quad (11.10)$$

where $\delta(x)$ is the Dirac delta function; that is, in this test case, h does not distribute. It is noteworthy that the resultant algorithm obtained based on the Gaussian prior can be applied to the case of the Laplace prior without modification [17].

11.3 Empirical Bayes Method

Using the empirical Bayes method, the values of the hyperparameters, H and γ , can be inferred from the observed dataset, \mathcal{D} . For the empirical Bayes method, a marginal log-likelihood function is defined as

$$L_{\text{EB}}(H, \gamma) := \frac{1}{nN} \ln [P(\mathcal{D} | h, \mathbf{J})]_{h, \mathbf{J}}, \quad (11.11)$$

where $[\cdot \cdot \cdot]_{h, \mathbf{J}}$ is the average over the prior distributions, that is,

$$[\cdot \cdot \cdot]_{h, \mathbf{J}} := \int d\mathbf{J} \int dh (\cdot \cdot \cdot) P_{\text{prior}}(h | H) P_{\text{prior}}(\mathbf{J} | \gamma).$$

This marginal log-likelihood function is referred to as the *empirical Bayes likelihood function* in this section. From the perspective of the empirical Bayes method, the optimal values of the hyperparameters, \hat{H} and $\hat{\gamma}$, are obtained by maximizing the empirical Bayes likelihood function, that is,

$$\{\hat{H}, \hat{\gamma}\} = \arg \max_{H, \gamma} L_{\text{EB}}(H, \gamma). \quad (11.12)$$

It is noteworthy that $[P(\mathcal{D} | h, \mathbf{J})]_{h, \mathbf{J}}$ in Eq. (11.11) is identified as the evidence appearing in Eq. (11.5).

The marginal log-likelihood function can be rewritten as

$$L_{\text{EB}}(H, \gamma) = \frac{1}{nN} \ln \left[\exp(nNL_{\text{ML}}(h, \mathbf{J})) \right]_{h, \mathbf{J}}. \quad (11.13)$$

Consider the case $N \gg n$. In this case, by using the saddle point evaluation, Eq. (11.13) is reduced to

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

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 $\{\hat{h}_{\text{ML}}, \hat{\mathbf{J}}_{\text{ML}}\}$ (i.e., the solution for BML) are inserted. This indicates that parameter estimations can be conducted independently of hyperparameter estimation. This trivial case is not considered in this section. Remember that the objective is to estimate the hyperparameter values *without* estimating the specific values of the parameters.

11.4 Statistical Mechanical Analysis of Empirical Bayes Likelihood

The empirical Bayes likelihood function in Eq. (11.11) involves intractable multiple integrations. This section presents an evaluation of the empirical Bayes likelihood function using statistical mechanical analysis. The outline of the evaluation is as follows. First, the intractable multiple integrations in Eq. (11.11) are evaluated using the *replica method* [18, 19]. This evaluation leads to a quantity with a certain intractable multiple summation. The quantity is approximately evaluated using the *Plefka expansion* [3]. Thus, from the two approximations, the replica method and Plefka expansion, the evaluation result for the empirical Bayes likelihood function is obtained.

11.4.1 Replica Method

The empirical Bayes likelihood function in Eq. (11.11) can be represented as

$$L_{\text{EB}}(H, \gamma) = \frac{1}{nN} \ln \lim_{x \rightarrow -1} \Psi_x(H, \gamma), \quad (11.14)$$

where

$$\Psi_x(H, \gamma) := \left[Z(h, \mathbf{J})^{xN} \exp N \left(h \sum_{i=1}^n d_i + \sum_{i < j} J_{ij} d_{ij} \right) \right]_{h, \mathbf{J}}, \quad (11.15)$$

and

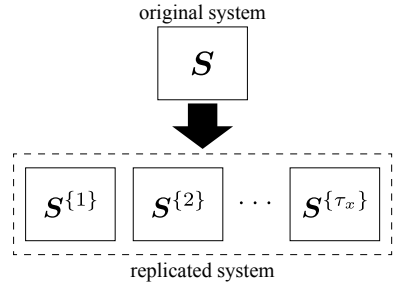
$$d_i := \frac{1}{N} \sum_{\mu=1}^N S_i^{(\mu)}, \quad d_{ij} := \frac{1}{N} \sum_{\mu=1}^N S_i^{(\mu)} S_j^{(\mu)}$$

are the sample averages of the observed data points. We now assume that $\tau_x := xN$ is a natural number larger than zero. Accordingly, Eq. (11.15) can be expressed as

$$\begin{aligned} \Psi_x(H, \gamma) = & \left[\sum_{\mathcal{S}_x} \exp \left\{ h \sum_{i=1}^n \left(\sum_{a=1}^{\tau_x} S_i^{(a)} + N d_i \right) \right. \right. \\ & \left. \left. + \sum_{i < j} J_{ij} \left(\sum_{a=1}^{\tau_x} S_i^{(a)} S_j^{(a)} + N d_{ij} \right) \right\} \right]_{h, \mathbf{J}}, \end{aligned} \quad (11.16)$$

where $a, b \in \{1, 2, \dots, \tau_x\}$ are the replica indices and $S_i^{(a)}$ is the i th variable in the a th replica. $\mathcal{S}_x := \{S_i^{(a)} \mid i = 1, 2, \dots, n; a = 1, 2, \dots, \tau_x\}$ is the set of all vari-

Fig. 11.2 Illustration of the replicated system. The τ_x replicas, $\mathcal{S}^{\{1\}}, \mathcal{S}^{\{2\}}, \dots, \mathcal{S}^{\{\tau_x\}}$, arise from $Z(h, \mathbf{J})^{\tau_x}$ in Eq. (11.15)



ables in the replicated system (see Fig. 11.2) and $\sum_{\mathcal{S}_x}$ is the sum over all possible configurations of \mathcal{S}_x , that is,

$$\sum_{\mathcal{S}_x} := \prod_{i=1}^n \prod_{a=1}^{\tau_x} \sum_{S_i^{[a]} = \pm 1} .$$

We evaluate $\Psi_x(H, \gamma)$ under the assumption that τ_x is a natural number, and then we take the limit of $x \rightarrow -1$ from the evaluation result as an analytic continuation,¹ to obtain the empirical Bayes likelihood function (this is the so-called *replica trick*).

By employing the Gaussian prior in Eq. (11.8), Eq. (11.16) becomes

$$\Psi_x^{\text{Gauss}}(H, \gamma) = \exp \left\{ nNHM + \frac{\gamma(n-1)N^2}{4} \left(C_2 + \frac{x}{N} \right) - F_x(H, \gamma) \right\}, \quad (11.17)$$

where

$$M := \frac{1}{n} \sum_{i=1}^n d_i, \quad C_k := \frac{2}{n(n-1)} \sum_{i < j} d_{ij}^k, \quad (11.18)$$

and

$$F_x(H, \gamma) := - \ln \sum_{\mathcal{S}_x} \exp \left(- E_x(\mathcal{S}_x; H, \gamma) \right) \quad (11.19)$$

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

¹ The justification for this analytic continuation may not be guaranteed mathematically. Thus, this type of analysis is regarded as “trick.”

$$\begin{aligned}
E_x(\mathcal{S}_x; H, \gamma) := & -H \sum_{i=1}^n \sum_{a=1}^{\tau_x} S_i^{[a]} - \frac{\gamma N}{n} \sum_{i < j} d_{ij} \sum_{a=1}^{\tau_x} S_i^{[a]} S_j^{[a]} \\
& - \frac{\gamma}{n} \sum_{i < j} \sum_{a < b} S_i^{[a]} S_j^{[a]} S_i^{[b]} S_j^{[b]}
\end{aligned} \tag{11.20}$$

is the Hamiltonian (or energy function) of the replicated system, where $\sum_{a < b}$ is the sum over all distinct pairs of replicas, that is, $\sum_{a < b} = \sum_{a=1}^{\tau_x} \sum_{b=a+1}^{\tau_x}$.

11.4.2 Plefka Expansion

Because the replicated free energy in Eq. (11.19) includes intractable multiple summations, an approximation is required to proceed with the current evaluation. In this section, the replicated free energy in Eq. (11.19) is approximated using the Plefka expansion [3]. In brief, the Plefka expansion is a perturbative expansion in Gibbs free energy that is a dual form of a corresponding Helmholtz free energy.

The Gibbs free energy is obtained as

$$G_x(m, H, \gamma) = -n\tau_x Hm + \text{extr}_{\lambda} \left\{ \lambda n\tau_x m - \ln \sum_{\mathcal{S}_x} \exp(-E_x(\mathcal{S}_x; \lambda, \gamma)) \right\}. \tag{11.21}$$

The derivation of this Gibbs free energy is described in Sect. 11.7.1. The summation in Eq. (11.21) can be performed when $\gamma = 0$, which gives

$$\begin{aligned}
G_x(m, H, 0) &= -n\tau_x Hm + n\tau_x \text{extr}_{\lambda} \left\{ \lambda m - \ln(2 \cosh \lambda) \right\} \\
&= -n\tau_x Hm + n\tau_x e(m),
\end{aligned} \tag{11.22}$$

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

$$e(m) := \frac{1+m}{2} \ln \frac{1+m}{2} + \frac{1-m}{2} \ln \frac{1-m}{2}. \tag{11.23}$$

In the context of the Plefka expansion, the Gibbs free energy $G_x(m, H, \gamma)$ is approximated by the perturbation from $G_x(m, H, 0)$. Expanding $G_x(m, H, \gamma)$ around $\gamma = 0$ gives

$$\frac{G_x(m, H, \gamma)}{nN} = -xHm + xe(m) + \phi_x^{(1)}(m)\gamma + \phi_x^{(2)}(m)\gamma^2 + O(\gamma^3), \tag{11.24}$$

where $\phi_x^{(1)}(m)$ and $\phi_x^{(2)}(m)$ are the expansion coefficients defined by

$$\phi_x^{(k)}(m) := \frac{1}{nNk!} \lim_{\gamma \rightarrow 0} \frac{\partial^k G_x(m, H, \gamma)}{\partial \gamma^k}.$$

The forms of the two coefficients are presented in Eqs. (11.34) and (11.35) in Sect. 11.7.2.

From Eqs. (11.14), (11.17), (11.24), and (11.33), the approximation of the empirical Bayes likelihood function is obtained as

$$L_{\text{EB}}(H, \gamma) \approx HM - \text{extr}_m \left[Hm - e(m) + \Phi(m)\gamma + \phi_{-1}^{(2)}(m)\gamma^2 \right], \quad (11.25)$$

where

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

The forms of $\phi_{-1}^{(1)}(m)$ and $\phi_{-1}^{(2)}(m)$ are presented in Eqs. (11.37) and (11.38) in Sect. 11.7.2.

11.4.3 Algorithm for Hyperparameter Estimation

As mentioned in Sect. 11.3, the empirical Bayes inference is achieved by maximizing $L_{\text{EB}}(H, \gamma)$ with respect to H and γ (cf. Eq. (11.12)). The extremum condition in Eq. (11.25) with respect to H leads to

$$\hat{m} = M, \quad (11.26)$$

where \hat{m} is the value of m that satisfies the extremum condition in Eq. (11.25). By combining the extremum condition of Eq. (11.25) with respect to m with Eq. (11.26),

$$\hat{H} = \text{atanh}M - \left(\frac{\partial \phi_{-1}^{(1)}(M)}{\partial M} \gamma + \frac{\partial \phi_{-1}^{(2)}(M)}{\partial M} \gamma^2 \right) \quad (11.27)$$

is obtained, where $\text{atanh}x$ is the inverse function of $\tanh x$. From Eqs. (11.25) and (11.26), the optimal value of γ is obtained by

$$\hat{\gamma} = \arg \max_{\gamma} \left[-\Phi(M)\gamma - \phi_{-1}^{(2)}(M)\gamma^2 \right]. \quad (11.28)$$

Since Eq. (11.28) represents a univariate quadratic optimization, $\hat{\gamma}$ is immediately obtained as follows: (i) when $\phi_{-1}^{(2)}(M) > 0$ and $\Phi(M) \geq 0$ 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) $\hat{\gamma} \rightarrow \infty$, elsewhere. The case of $\phi_{-1}^{(2)}(M) = \Phi(M) = 0$ is ignored because it may be rarely observed in realistic settings. Using Eqs. (11.27)

and (11.28), the solution to the empirical Bayes inference can be obtained without any iterative process. The pseudocode of the presented procedure is shown in Algorithm 1. The order of the computational complexity of the presented method is $O(Nn^2)$. Remember that the order of the computational complexity of the exact ML estimation is $O(2^n)$.

Algorithm 1 Proposed Inference Algorithm

- 1: **Input** Observed dataset: $\mathcal{D} := \{\mathbf{S}^{(\mu)} \in \{-1, +1\}^n \mid \mu = 1, 2, \dots, N\}$.
- 2: Compute M , Ω , C_1 , and C_2 using the dataset according to Eqs. (11.18) and (11.36).
- 3: Determine $\hat{\gamma}$ using Eq. (11.28):

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

where case (i): $\phi_{-1}^{(2)}(M) > 0$, $\Phi(M) \geq 0$ 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 Eq. (11.27).
 - 5: **Output** $\hat{\gamma}$ and \hat{H} .
-

In the presented method, the value of \hat{H} does not affect the determination of $\hat{\gamma}$. Several mean-field-based methods for BML (e.g., listed in Sect. 11.1) have similar procedures, in which $\hat{\mathbf{J}}_{\text{ML}}$ is determined separately from \hat{h}_{ML} . This is a common property of the mean-field-based methods for BML, including the current empirical Bayes problem.

Although the presented method is derived based on the Gauss prior presented in Eq. (11.8), the same procedure can be applied to the case of the Laplace prior presented in Eq. (11.9) [17].

11.5 Demonstration

This section discusses the results of numerical experiments. In these experiments, the observed dataset \mathcal{D} was generated by the generative Boltzmann machine (gBM), which has the same form as Eq. (11.1), via Gibbs sampling (with a simulated-annealing-like strategy). The parameters of gBM were drawn from the prior distributions in Eqs. (11.4) and (11.10). This implies that the model-matched case (i.e., the generative and learning models are identical) was considered. In the following, the notation $\alpha := N/n$ and $J := \sqrt{\gamma}$ are used. The standard deviations of the Gaussian prior in Eq. (11.8) and the Laplace prior in Eq. (11.9) can thus be represented as J/\sqrt{n} . The hyperparameters of gBM are denoted by H_{true} and J_{true} .

11.5.1 Gaussian Prior Case

We now consider the case in which the prior distribution of \mathbf{J} is the Gaussian prior in Eq. (11.8). In this case, the Boltzmann machine corresponds to the Sherrington-Kirkpatrick (SK) model [24], and thus exhibits a spin-glass transition at $J = 1$ when $h = 0$ (i.e., when $H = 0$).

We consider the case $H_{\text{true}} = 0$. The scatter plots for the estimation of \hat{J} for various J_{true} when $H_{\text{true}} = 0$ and $\alpha = 0.4$ are shown in Fig. 11.3. When $J_{\text{true}} < 1$, our estimates of \hat{J} are significantly consistent 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 several mean-field approximations. Figure 11.4 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. In our experiments, at least, the optimal value of α seems to be $\alpha_{\text{opt}} \approx 0.4$ when $H_{\text{true}} = 0$. Our method can also estimate \hat{H} . The results for the estimation of \hat{H} when $H_{\text{true}} = 0$ and $\alpha = 0.4$ are shown in Fig. 11.5. Figure 11.5a, b shows 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 where $J_{\text{true}} > 1$.

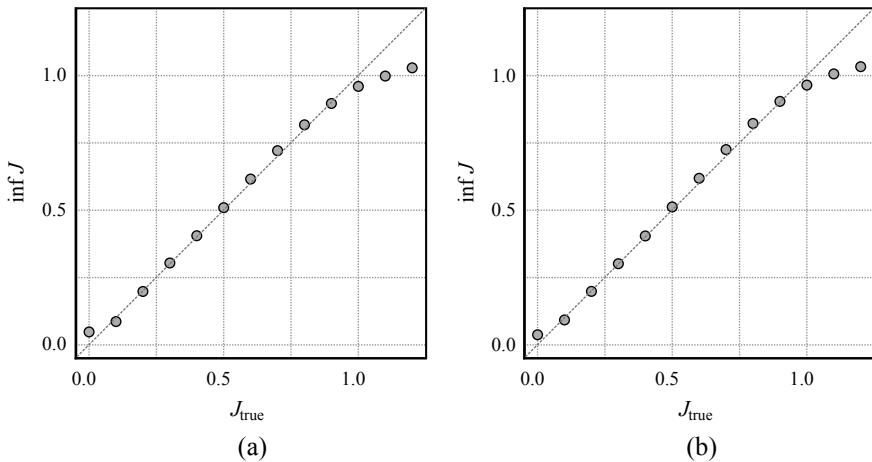


Fig. 11.3 Scatter plots of J_{true} (horizontal axis) versus \hat{J} (vertical axis) when $H_{\text{true}} = 0$ and $\alpha = 0.4$: **a** $n = 300$ and **b** $n = 500$. These plots represent the average values over 300 experiments

11.5.2 Laplace Prior Case

We now consider the case in which the prior distribution of \mathbf{J} is the Laplace prior in Eq. (11.9). The scatter plots for the estimation of \hat{J} for various values of J_{true} when $H_{\text{true}} = 0$ are shown in Fig. 11.6. The plots shown in Fig. 11.6 almost completely overlap with those in Fig. 11.4.

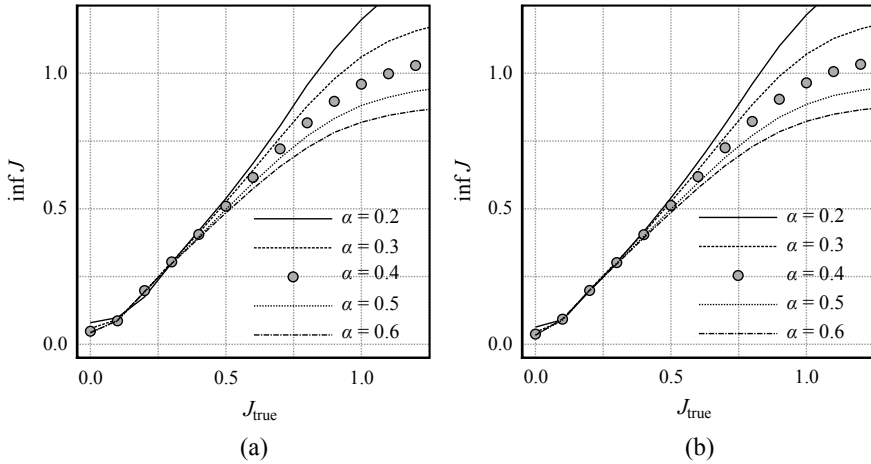


Fig. 11.4 Scatter plots of J_{true} (horizontal axis) versus \hat{J} (vertical axis) for various $\alpha = N/n$ when $H_{\text{true}} = 0$: **a** $n = 300$ and **b** $n = 500$. These plots represent the average values over 300 experiments

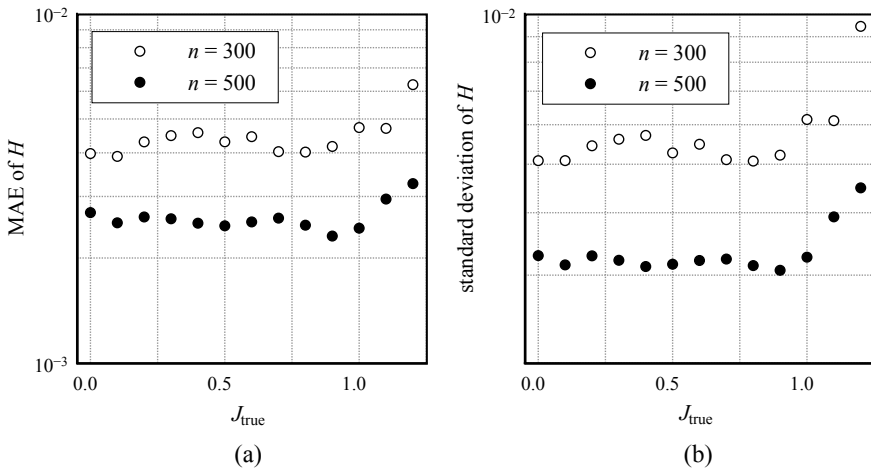


Fig. 11.5 Results of estimation of \hat{H} versus J_{true} when $H_{\text{true}} = 0$ and $\alpha = 0.4$: **a** the MAE and **b** standard deviation. These plots represent the average values over 300 experiments

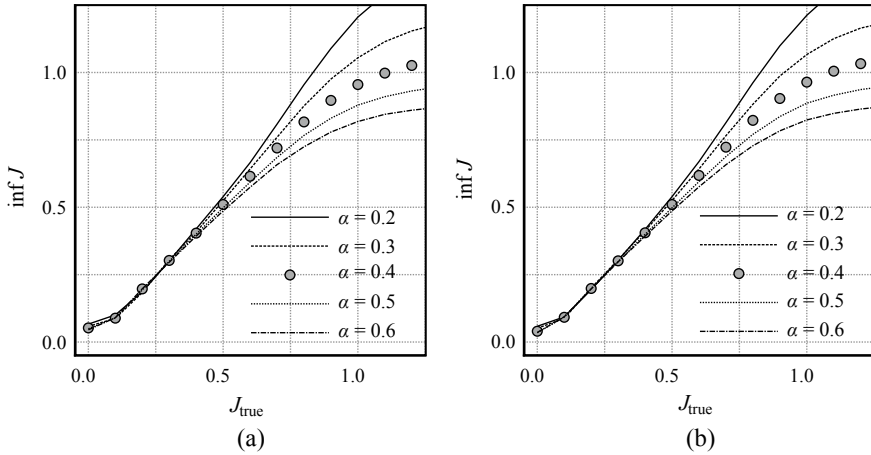


Fig. 11.6 Scatter plots of J_{true} (horizontal axis) versus \hat{J} (vertical axis) for various $\alpha = N/n$, when $H_{\text{true}} = 0$, in the case of the Laplace prior: **a** $n = 300$ and **b** $n = 500$. These plots represent the average values over 300 experiments

11.6 Summary and Discussion

This chapter describes the hyperparameter inference algorithm proposed in [17]. As evident from the numerical experiments, the proposed inference method in both the Gaussian and Laplace prior cases works efficiently except for the spin-glass phase. However, the presented method has the drawback that it is sensitive to the value of $\alpha = N/n$. In the experiments in Sect. 11.5, although $\alpha \approx 0.4$ was appropriate when $H_{\text{true}} = 0$, it is known that the appropriate value decreases as H_{true} increases [17]. Since we cannot know the value of H_{true} in advance, the appropriate setting of α is also unknown. Estimation of α_{opt} is an open problem. It seems to be unnatural that there exists an optimal value of α because larger datasets are better in usual machine learning. Such peculiar behavior can be attributed to the truncating approximation in the Pfleka expansion. A more detailed discussion of this issue is presented in [17].

Finally, we review the presented method from the perspective of sublinear computation without considering the aforementioned issues. The Boltzmann machine given in Eq. (11.1) has p parameters, where $p = O(n^2)$. In usual machine learning, $N = O(p)$ is, at least, required to obtain a good ML estimate for the Boltzmann machine. Therefore, a hyperparameter inference “without” the empirical Bayes method (namely, the strategy in which the hyperparameters are inferred through the ML estimate in a similar manner as that discussed in the latter part of Sect. 11.3) requires a dataset of size $O(p)$. However, the presented method requires only $N = O(n) = O(\sqrt{p})$ because $\alpha = O(1)$ with respect to n .

Acknowledgements This work was partially supported by JSPS KAKENHI (Grant Numbers: 15H03699, 18K11459, 18H03303, 25120013, and 17H00764), JST CREST (Grant Number: JPMJCR1402), and the COI Program from the JST (Grant Number JPMJCE1312).

11.7 Appendices

11.7.1 Appendix 1: Gibbs Free Energy

In this appendix, we derive the Gibbs free energy for the replicated (Helmholtz) free energy in Eq. (11.19).

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

$$f[Q] := \sum_{\mathcal{S}_x} E_x(\mathcal{S}; H, \gamma) Q(\mathcal{S}_x) + \sum_{\mathcal{S}_x} Q(\mathcal{S}_x) \ln Q(\mathcal{S}_x), \quad (11.29)$$

under the normalization constraint, that is, $\sum_{\mathcal{S}_x} Q(\mathcal{S}_x) = 1$, 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 Eq. (11.20).

The Gibbs free energy is obtained by adding new constraints to the minimization of $f[Q]$. We add the relationship

$$m = \frac{1}{n\tau_x} \sum_{i=1}^n \sum_{a=1}^{\tau_x} \sum_{\mathcal{S}_x} S_i^{[a]} Q(\mathcal{S}_x) \quad (11.30)$$

as the constraint. Using Lagrange multipliers, the Gibbs free energy is obtained as

$$G_x(m, H, \gamma) := \text{extr}_{Q, \lambda, r} \left\{ f[Q] - r \left(\sum_{\mathcal{S}_x} Q(\mathcal{S}_x) - 1 \right) - \lambda \left(\sum_{i=1}^n \sum_{a=1}^{\tau_x} \sum_{\mathcal{S}_x} S_i^{[a]} Q(\mathcal{S}_x) - n\tau_x m \right) \right\}, \quad (11.31)$$

where “extr” denotes the extremum with respect to the assigned parameters, and r and λ are the Lagrange multipliers for the normalization constraint of $Q(\mathcal{S}_x)$ and the constraint in Eq. (11.30), respectively. Performing the extremum operation with respect to $Q(\mathcal{S})$ and r in Eq. (11.31) gives

$$G_x(m, H, \gamma) = \text{extr}_{\lambda} \left\{ \lambda n \tau_x m - \ln \sum_{\mathcal{S}_x} \exp(-E_x(\mathcal{S}_x; H + \lambda, \gamma)) \right\}. \quad (11.32)$$

The replicated free energy in Eq. (11.19) coincides with the extremum of this Gibbs free energy with respect to m , that is,

$$F_x(H, \gamma) = \text{extr}_m G_x(m, H, \gamma). \quad (11.33)$$

By performing the shift $H + \lambda \rightarrow \lambda$ in Eq. (11.32), Eq. (11.21) is obtained.

11.7.2 Appendix 2: Coefficients of Plefka Expansion

This appendix presents the coefficients of the Plefka expansion in Eq. (11.24). Refer to Ref. [17] for a detailed derivation. The first-order coefficient is given by

$$\phi_x^{(1)}(m) = -\frac{x(n-1)NC_1}{2n}m^2 - \frac{(n-1)K_x}{2nN}m^4, \quad (11.34)$$

where $K_x := \tau_x(\tau_x - 1)/2$. The second-order coefficient is given by

$$\begin{aligned} \phi_x^{(2)}(m) = & -\frac{(n-1)^2\tau_x N\Omega}{2n^2}m^2(1-m^2) - \frac{(n-1)\tau_x NC_2}{4n^2}(1-m^2)^2 \\ & - \frac{(n-1)K_x C_1}{n^2}m^2(1-m^2)^2 - \frac{(n-1)K_x}{2n^2N}(n+\tau_x-3)m^4(1-m^2)^2 \\ & - \frac{(n-1)K_x}{4n^2N}(1-m^4)^2, \end{aligned} \quad (11.35)$$

where Ω in the first term of Eq. (11.35) is defined as

$$\Omega := \frac{1}{n} \sum_{i=1}^n \omega_i^2, \quad \omega_i := \frac{1}{n-1} \sum_{j \in \partial(i)} d_{ij} - C_1, \quad (11.36)$$

where $\partial(i) := \{1, 2, \dots, n\} \setminus \{i\}$. When $x = -1$, these coefficients are

$$\begin{aligned} \phi_{-1}^{(1)}(m) = & \frac{(n-1)NC_1}{2n}m^2 - \frac{(n-1)(N+1)}{4n}m^4, \quad (11.37) \\ \phi_{-1}^{(2)}(m) = & \frac{(n-1)^2N^2\Omega}{2n^2}m^2(1-m^2) + \frac{(n-1)N^2C_2}{4n^2}(1-m^2)^2 \\ & - \frac{(n-1)N(N+1)C_1}{2n^2}m^2(1-m^2)^2 \\ & - \frac{(n-1)(N+1)}{4n^2}(n-N-3)m^4(1-m^2)^2 - \frac{(n-1)(N+1)}{8n^2}(1-m^4)^2. \end{aligned} \quad (11.38)$$

References

1. D.H. Ackley, G.E. Hinton, T.J. Sejnowski, A learning algorithm for Boltzmann machines. *Cognit. Sci.* **9**, 147–169 (1985)
2. Y. Roudi, E. Aurell, J. Hertz, Statistical physics of pairwise probability models. *Front. Comput. Neurosci.* **3**, 1–22 (2009)
3. T. Plefka, Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J. Phys. A Math. Gen.* **15**(6), 1971–1978 (1982)
4. A. Pelizzola, Cluster variation method in statistical physics and probabilistic graphical models. *J. Phys. A Math. Gen.* **38**(33), R309 (2005)
5. H.J. Kappen, F.B. Rodríguez, Efficient learning in Boltzmann machines using linear response theory. *Neural Comput.* **10**(5), 1137–1156 (1998)
6. T. Tanaka, Mean-field theory of Boltzmann machine learning. *Phys. Rev. E* **58**, 2302–2310 (1998)
7. M. Yasuda, T. Horiguchi, Triangular approximation for information ising model and its application to Boltzmann machine. *Physica A* **368**, 83–95 (2006)
8. V. Sessak, R. Monasson, Small-correlation expansions for the inverse Ising problem. *J. Phys. A Math. Theoret.* **42**(5) (2009)
9. M. Yasuda, K. Tanaka, Approximate learning algorithm in Boltzmann machines. *Neural Comput.* **21**(11), 3130–3178 (2009)
10. F. Ricci-Tersenghi, The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *J. Stat. Mech. Theory Experi.* **2012**(08), P08015 (2012)
11. C. Furtlehner, Approximate inverse Ising models close to a Bethe reference point. *J. Stat. Mech. Theor. Exp.* **2013**(09), P09020 (2013)
12. J. Sohl-Dickstein, P.B. Battaglino, M.R. DeWeese, New method for parameter estimation in probabilistic models: minimum probability flow. *Phys. Rev. Lett.* **107** (2011)
13. M. Yasuda, Monte Carlo integration using spatial structure of Markov random field. *J. Phys. Soc. Jpn.* **84**(3) (2015)
14. M. Yasuda, K. Uchizawa, A generalization of spatial monte carlo integration. *Neural Comput.* **33**(4), 1037–1062 (2021)
15. D.J.C. MacKay, Bayesian interpolation. *Neural Comput.* **4**(3), 415–447 (1992)
16. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006)
17. M. Yasuda, T. Obuchi, Empirical Bayes method for Boltzmann machines. *J. Phys. A Math. Theoret.* **53**(1), 014004 (2019)
18. M. Mezard, G. Parisi, M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, Singapore, 1987)
19. H. Nishimori, *Statistical Physics of Spin Glass and Information Processing—Introduction* (Oxford University Press, 2001)
20. T. Rizzo, A. Lage-Castellanos, R. Mulet, F. Ricci-Tersenghi, Replica cluster variational method. *J. Stat. Phys.* **139**, 375–416 (2010)
21. M. Yasuda, Y. Kabashima, K. Tanaka, Replica plefka expansion of Ising systems. *J. Stat. Mech. Theor. Exp.* P04002 (2012)
22. A. Lage-Castellanos, R. Mulet, F. Ricci-Tersenghi, T. Rizzo, Replica cluster variational method: the replica symmetric solution for the 2d random bond ising model. *J. Phys. A Math. Theor.* **46**(13) (2013)
23. M. Yasuda, S. Kataoka, K. Tanaka, Statistical analysis of loopy belief propagation in random fields. *Phys. Rev. E* **92**, 042120 (2015)
24. D. Sherrington, S. Kirkpatrick, Solvable model of a spin-glass. *Phys. Rev. Lett.* **35**, 1792–1796 (1975)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

