Dynamics of the maximum marginal likelihood hyperparameter estimation in image restoration: Gradient descent versus expectation and maximization algorithm

Jun-ichi Inoue
Complex Systems Engineering, Graduate School of Engineering, Hokkaido University, N13-W8, Kita-ku, Sapporo 060-8628, Japan

Kazuyuki Tanaka
Department of Computer and Mathematical Sciences, Graduate School of Information Sciences, Tohoku University, Aramaki-aza-aoba 04, Aoba-ku, Sendai 980-8579, Japan

(Received 6 July 2001; published 19 December 2001)

Dynamical properties of image restoration and hyperparameter estimation are investigated by means of statistical mechanics. We introduce an exactly solvable model for image restoration and derive differential equations with respect to macroscopic quantities. From these equations, we evaluate relaxation processes of the system to the equilibrium state. Our statistical mechanical approach also enables us to investigate the hyperparameter estimation by means of maximization of the marginal likelihood by using gradient descent and the expectation and maximization algorithm from the dynamical point of view.

DOI: 10.1103/PhysRevE.65.016125 PACS number(s): 02.50.−r, 05.20.−y

I. INTRODUCTION

As a typical massive system, image restoration based on the Markov random field (MRF) model has been investigated by the statistical mechanical technique of disordered spin systems [1–4]. Among these results, statistical mechanical analysis succeeded in evaluating the measure of success for image restoration and made the hyperparameter dependence clear [2–4]. However, all of that research was restricted to studies of static properties of image restoration. In the context of the Bayesian statistical approach, we usually use the Markov chain Monte Carlo (MCMC) method to obtain a *maximum a posteriori* (MAP) estimate by simulated annealing [5], or to calculate expectations over posterior distribution for *maximum posterior marginal* (MPM) estimation [6]. In the recent study by Nishimori and Wong [2], they introduced an infinite range mean-field version of the MRF model and calculated the overlap between the original image and restored one analytically. However, they did not investigate the dynamical process of image restoration, that is to say, the process of the MCMC method by Glauber dynamics to obtain the MPM estimate. Although it is worthwhile to investigate such dynamical processes in image restoration, relatively little progress has been made in the theoretical understanding of them. Recently, Inoue and Carlucci [4] investigated dynamical properties of gray-scale image restoration using the mean-field Q-Ising spin glass model analytically. They found that the MPM estimate gets worse than the degraded image when one fails to set the hyperparameters appropriately. Therefore, it is important to study how we should infer the optimal hyperparameters. As an approach to estimate the optimal hyperparameters, the *maximum marginal likelihood* (MML) method has been used by many authors in practical situations [7,18]. If one maximizes the marginal likelihood by gradient descent, Boltzmann machine-type learning equations are obtained and these equations contain expectations over both posterior and prior distributions. In order to carry out those expectations, we usually use the MCMC method. However, it is hard to evaluate the performance of the MML estimation due to difficulties in simulating the thermodynamically equilibrium state within reliable precision. Therefore, we need some analytical and rigorous studies on the hyperparameter estimation. Obviously, the learning process of the hyperparameter estimation and the stochastic process of the MCMC method as *dynamics*. From the viewpoint of statistical mechanics of spin systems, the process of the hyperparameter estimation is regarded as a dynamics of the spin system in which coupling constant and field strength are time-dependent variables. Then, the time dependence of these variables is determined by the algorithm we choose to maximize the marginal likelihood. As far as we know, no studies have ever tried to investigate those dynamical properties analytically. In this paper, we investigate dynamical properties of image restoration including hyperparameter estimation by using the statistical mechanical technique.

This paper is organized as follows. In Sec. II, according to Nishimori and Wong [2], we explain statistical mechanical formulation of image restoration in the context of the MPM estimation. In Sec. III we derive differential equations with respect to macroscopic observables of the infinite range mean-field MRF model from the microscopic Master equation. By solving these differential equations, we discuss the relaxation process of image restoration. In Sec. IV marginal likelihood as a function of hyperparameters is calculated by the replica method. We also derive Boltzmann machine-type learning equations to maximize the marginal likelihood by gradient descent. Flows in hyperparameter space are obtained by analyzing the learning equations. In the same section, we investigate the performance of the EM (expectation and maximization) algorithm [8] which is widely used to estimate hyperparameters from incomplete data sets. It is well known that the EM algorithm shows faster convergence at the beginning of the algorithm than some other algorithm does. However, there is no study to make this property clear by using some solvable models. In this section we compare the performance of the EM algorithm with that of gradient descent explicitly. Section V is devoted to the summary.

II. STATISTICAL MECHANICAL FORMULATION FOR IMAGE RESTORATION

In this section we explain how we formulate image restoration as a problem of a disordered spin system. According to Nishimori and Wong [2], we consider a black and while image. Then, an original image is denoted by an N-dimensional vector $\{\xi\}\equiv(\xi_1,\xi_2,\ldots,\xi_N)$ and each pixel ξ_i takes ± 1. These pixels are located on an arbitrary lattice in two dimension. In order to treat image restoration by statistical mechanics of disordered spin systems, we should assume that the original image is given by *a priori* Boltzmann-Gibbs distribution

$$P(\{\xi\})=\frac{\exp\left(\beta_s\sum_{ij}\ \xi_i\xi_j\right)}{Z_s},\quad Z_s=\sum_{\xi}\ \exp\left(\beta_s\sum_{ij}\ \xi_i\xi_j\right), \tag{1}$$

where $\Sigma_{ij}(\cdots)$ is carried out for all nearest neighboring pixels. Thus, we use a snapshot of the MCMC simulation for the ferromagnetic Ising model as an original image. $T_s(\equiv\beta_s^{-1})$ appearing in the argument of the exponential (1) corresponds to temperature. We obtain pictures of all black or all white when we set $T_s\to 0$, while we obtain random noise pictures in the limit of $T_s\to\infty$. A particular original image $\{\xi\}$ is degraded to a particular damaged picture $\{\tau\}\equiv(\tau_1,\tau_2,\cdots,\tau_N)$ by a noise channel represented by the following conditional probability:

$$P(\{\tau\}|\{\xi\})=\frac{\exp\left(\beta_\tau\sum_{i}\ \tau_i\xi_i\right)}{(2\cosh\beta_\tau)^N}, \tag{2}$$

where the sum $\Sigma_i(\cdots)$ is carried out for all pixels and we assumed that each pixel is degraded independently. β_τ represents a noise level of the channel because the above expression is rewritten as $P(-\xi_i|\xi_i)=p=1-P(\xi_i|\xi_i)$ with $p=e^{-\beta_\tau}/(e^{\beta_\tau}+e^{-\beta_\tau})$ for all pixels independently. Therefore, this kind of noise is referred to as the *binary symmetric channel* (BSC).

The BSC is easily extended to the *Gaussian channel* (GC) as follows:

$$P(\{\tau\}|\{\xi\})=\frac{1}{(\sqrt{2\pi}\tau)^N}\exp\left(-\frac{\sum_i\ (\tau_i-\tau_0\xi_i)^2}{2\tau^2}\right)$$
$$=F_{\mathrm{GC}}(\{\tau\})\exp\left(\frac{\tau_0}{\tau^2}\sum_i\ \tau_i\xi_i\right), \tag{3}$$

$$F_{\mathrm{GC}}(\{\tau\})\equiv\frac{1}{(\sqrt{2\pi}\tau)^N}\exp\left(-\frac{\sum_i\ (\tau_i^2+\tau_0^2)}{2\tau^2}\right). \tag{4}$$

If we replace $F_{\mathrm{GC}}(\{\tau\})$ appearing in Eq. (3) by

$$F_{\mathrm{BSC}}(\{\tau\})\equiv\frac{1}{(2\cosh\beta_\tau)^N}\prod_i\ \{\delta(\tau_i-1)+\delta(\tau_i+1)\} \tag{5}$$

with $\tau_0/\tau^2=\beta_\tau$, the BSC [Eq. (2)] is recovered. We should notice that a sum $\Sigma_\tau\Omega(\{\tau\})$ for an arbitrary function $\Omega(\{\tau\})$ is calculated in terms of $F_{\mathrm{GC,BSC}}(\{\tau\})$ as

$$\sum_\tau\ \Omega(\{\tau\})=\int\cdots\int d\{\tau\}F_{\mathrm{GC,BSC}}(\{\tau\})\Omega(\{\tau\}), \tag{6}$$

where we defined $d\{\tau\}\equiv d\tau_1 d\tau_2\cdots d\tau_N$. Then, Bayes theorem gives the posterior distribution

$$P(\{\sigma\}|\{\tau\})=\frac{P(\{\tau\}|\{\sigma\})P(\{\sigma\})}{\sum_\sigma\ P(\{\tau\}|\{\sigma\})P(\{\sigma\})}$$
$$=\frac{e^{J\Sigma_{ij}\sigma_i\sigma_j+h\Sigma_i\tau_i\sigma_i}}{\sum_\sigma\ e^{J\Sigma_{ij}\sigma_i\sigma_j+h\Sigma_i\tau_i\sigma_i}}, \tag{7}$$

where J and h are hyperparameters and we introduced models of the prior [Eq. (1)] and the likelihood [Eq. (2)] as

$$P(\{\sigma\})=\frac{\exp\left(J\sum_{ij}\ \sigma_i\sigma_j\right)}{Z_\Pi},$$

$$P(\{\tau\}|\{\sigma\})=\frac{\exp\left(h\sum_i\ \tau_i\sigma_i\right)}{Z_L}, \tag{8}$$

respectively. A configuration $\{\sigma\}\equiv(\sigma_1,\sigma_2,\ldots,\sigma_N)$ denotes an estimate of a particular original image $\{\xi\}$. Z_Π and Z_L in Eq. (8) are normalization constants given by

$$Z_\Pi=\sum_\sigma\ \exp\left(J\sum_{ij}\ \sigma_i\sigma_j\right),\quad Z_L=\sum_\tau\ \exp\left(h\sum_i\ \tau_i\sigma_i\right). \tag{9}$$

It is important for us to bear in mind that Z_L is independent of $\{\sigma\}$ for both the BSC and the GC. Actually, Z_L leads to

$$Z_L=\int\cdots\int d\{\tau\}F_{\mathrm{BSC}}(\{\tau\})\exp\left(h\sum_i\ \tau_i\sigma_i\right)$$
$$=\left(\frac{2\cosh h}{2\cosh\beta_\tau}\right)^N \tag{10}$$

for the BSC and

$$Z_L = \int \cdots \int d\{\tau\} F_{\mathrm{GC}}(\{\tau\}) \exp\left(h \sum_i \tau_i \sigma_i \right)$$
$$= \exp\left(-\frac{N\tau_0^2}{2\tau^2} + \frac{N\tau^2 h^2}{2} \right) \tag{11}$$

for the GC.

In the context of MAP estimation, we choose the estimate $\{\sigma\}$ as a grand state of the following Hamiltonian (cost function):

$$\mathcal{H}(\{\sigma\}) = -J \sum_{ij} \sigma_i \sigma_j - h \sum_i \tau_i \sigma_i . \tag{12}$$

In order to obtain the grand state, we usually use simulated annealing [9] or mean field annealing [10].

On the other hand, in the context of MPM estimation, we first calculate the marginal distribution around a single pixel σ_i:

$$P(\sigma_i | \{\tau\}) = \sum_{\{\sigma\} \neq \sigma_i} P(\{\sigma\} | \{\tau\}) \tag{13}$$

and we choose the sign of the difference between $P(\sigma_i = +1 | \{\tau\})$ and $P(\sigma = -1 | \{\tau\})$ as an estimate of the ith pixel $\hat{\xi}_i$ as

$$\hat{\xi}_i = \mathrm{argmax}_{\sigma_i} P(\sigma_i | \{\tau\}) = \mathrm{sgn}\left(\sum_{\sigma_i = \pm 1} P(\sigma_i | \{\tau\}) \right)$$
$$= \mathrm{sgn}\left(\frac{\sum_\sigma \sigma_i P(\{\sigma\} | \{\tau\})}{\sum_\sigma P(\{\sigma\} | \{\tau\})} \right) \equiv \mathrm{sgn}(\langle \sigma_i \rangle_{J,h}). \tag{14}$$

In this expression, we defined $\langle \sigma_i \rangle_{J,h}$ as an average of the ith pixel σ_i over the posterior distribution (7) and this is written explicitly as

$$\langle \sigma_i \rangle_{J,h} = \frac{\sum_\sigma \sigma_i e^{J \Sigma_{ij} \sigma_i \sigma_j + h \Sigma_i \tau_i \sigma_i}}{\sum_\sigma e^{J \Sigma_{ij} \sigma_i \sigma_j + h \Sigma_i \tau_i \sigma_i}} . \tag{15}$$

This corresponds to a local magnetization of the spin system that is described by the Hamiltonian $\mathcal{H}(\{\sigma\})$ at temperature $T=1$. Thus, in order to investigate properties of the MPM estimation for image restoration, we should study the random field Ising model described by $\mathcal{H}(\{\sigma\})$. Then, we are interested in the quantity

$$M(J,h) \equiv \sum_{\xi,\tau} P(\{\xi\}) P(\{\tau\} | \{\xi\}) \xi_i \mathrm{sgn}(\langle \sigma_i \rangle_{J,h}), \tag{16}$$

which means the averaged overlap between an arbitrary original pixel ξ_i and the MPM estimate $\hat{\xi}_i = \mathrm{sgn}(\langle \sigma_i \rangle_{J,h})$. Apparently, the best restoration of the original image is achieved when the overlap M is as close to 1 as possible. For this averaged overlap $M(J,h)$, the next inequality holds [2],

$$M(J,h) \leqslant M(\beta_s, \beta_\tau). \tag{17}$$

This inequality means that the averaged overlap M takes its maximum when one sets the hyperparameters to their true values, namely, $J = \beta_s$ and $h = \beta_\tau$. However, it is impossible to derive the hyperparameter dependence of the overlap around its optimal value $M(\beta_s, \beta_\tau)$ from the above inequality. To investigate this dependence, Nishimori and Wong [2] introduced a mean-field version of the MRF model and calculated the overlap as a function of J and h. The mean-field model is rather an artificial model in which every pixel is connected to the others; however, this model is very useful to discuss the behavior of macroscopic quantities of the system, like the overlap M. Using the replica method [11], one obtains saddle point equations

$$m_0 \equiv \frac{1}{N} \sum_i \xi_i = \tanh(\beta_s m_0), \tag{18}$$

$$m \equiv \frac{1}{N} \sum_i \sigma_i = \frac{\sum_\xi e^{\beta_s m_0 \xi}}{2 \cosh(\beta_s m_0)} \int_{-\infty}^{\infty} Dx \tanh(Jm + \tau h x + \tau_0 h \xi), \tag{19}$$

$$M \equiv \frac{1}{N} \sum_i \xi_i \hat{\xi}_i = \frac{\sum_\xi e^{\beta_s m_0 \xi}}{2 \cosh(\beta_s m_0)} \int_{-\infty}^{\infty} Dx \xi \, \mathrm{sgn}(Jm + \tau h x + \tau_0 h \xi), \tag{20}$$

where we defined the Gaussian integral measure by $Dx \equiv dx\, e^{-x^2/2}/\sqrt{2\pi}$. Equation (18) determines macroscopic properties of the original image given by the Hamiltonian $-\Sigma_{ij} \xi_i \xi_j$ at temperature $T_s (\equiv \beta_s^{-1})$. From a statistical mechanical point of view, m_0 corresponds to the magnetization of the mean-field ferromagnetic Ising model. For a given T_s, one obtains m_0 by solving Eq. (18). Substituting T_s, m_0, and noise parameters τ_0 (a center of Gaussian) and τ (a standard deviation) into Eq. (19), one obtains magnetization m for the restored image system $\{\sigma\}$ as a function of $T_m (\equiv J^{-1})$ and h. Then, one substitutes $m(T_m, h)$ into the expression of M, and finds the hyperparameter dependence of the overlap explicitly. In Fig. 1 we plot the overlap M as a function of $1/J$ $(\equiv T_m)$. We set $\tau = \tau_0 = 1$ $(\beta_\tau = \tau_0/\tau^2 = 1)$ and temperature of the original image is chosen as $T_s = 0.9$. The overlap for the two cases of the field h, namely, $h = \beta_\tau T_s J = \tau_0 T_s J/\tau^2 = 0.9J \equiv h_{\mathrm{opt}}$ (a) and $h=1$ (b) are shown. We should notice that the MAP estimate is obtained in the limit of $T_m \to 0$ keeping the ratio h/J constant. Therefore, the overlap for the MAP estimate depends on the ratio h/J and takes its maximum when we set $h/J = \beta_\tau T_s = 0.9$ [see Fig. 1 (a)]. From this figure we see that the overlap takes its maximum at T_m

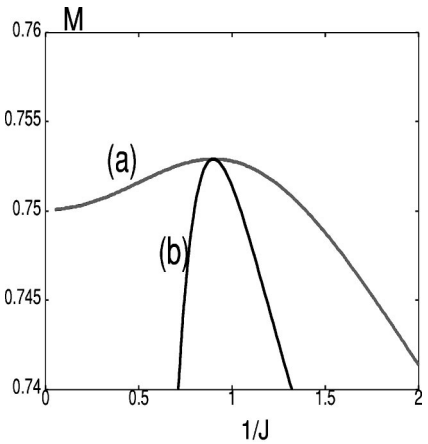

FIG. 1. $1/J(\equiv T_m)$ dependence of the overlap M. The temperature of the original image is $T_s=0.9$ and the noise level is $\beta_\tau=\tau_0/\tau^2=1(\tau_0=\tau=1)$. We set the field h as $h=\beta_\tau T_s J=(\tau_0 T_s/\tau^2)J=0.9J\equiv h_{\mathrm{opt}}$ (a) and $h=1$ (b). In the limit of $1/J\to 0$, we obtain the overlap of the MAP estimation. In both cases (a) and (b), the overlap M takes its maximum at $T_m=T_s=0.9$.

$=T_s=0.9$ and $h=\beta_\tau=\tau_0/\tau^2=1$. In the next section, we focus our attention on the dynamics of the MPM estimation.

III. DYNAMICS OF IMAGE RESTORATION

In the preceding section we showed the performance of the MPM estimation. However, in those calculations we assumed that the system already reached the equilibrium state. In other words, each state $\{\sigma\}$ obeys the Boltzmann-Gibbs distribution $\sim e^{-\mathcal{H}(\{\sigma\})}$. When we need to generate the distribution to calculate the MPM estimate $\mathrm{sgn}(\langle\sigma_i\rangle_{J,h})$, we often use the MCMC method and simulate the equilibrium states on computer. Therefore, it is important to study how the system relaxes to its equilibrium state and grasp the behavior of time evolutionary observables analytically. As far as we know, there is no research to deal with dynamics of image restoration including hyperparameter estimation analytically. In this section, for the infinite range mean-field MRF model, we derive differential equations with respect to macroscopic order parameters of the restored image system from the microscopic master equation.

First of all, we should remember that a transition rate $w_k(\{\sigma\})$ from $\{\sigma\}\equiv(\sigma_1,\sigma_2,\ldots,\sigma_k,\ldots,\sigma_N)$ to $\{\sigma\}'\equiv(\sigma_1,\sigma_2,\ldots,-\sigma_k,\ldots,\sigma_N)$ leads to

$$w_k(\{\sigma\})=\frac{1}{2}\{1-\sigma_k\tanh[h_k(\{\sigma\})]\},$$

$$h_k(\{\sigma\})=\frac{J}{N}\sum_j \sigma_j+h\tau_k \tag{21}$$

in the context of the Glauber dynamics of the MCMC method. It is important for us to bear in mind that the Hamiltonian $\mathcal{H}(\{\sigma\})$ of the system is rewritten in terms of $h_k(\{\sigma\})$ as

$$\mathcal{H}(\{\sigma\})=-\sum_k h_k(\{\sigma\})\sigma_k, \tag{22}$$

where we rescaled the coupling J as J/N to take a proper thermodynamic limit (the Hamiltonian should be of order N).

Then, probability $p_t(\{\sigma\})$ that the system visits a state $\{\sigma\}$ at time t obeys the master equation

$$\frac{dp_t(\{\sigma\})}{dt}=\sum_{k=1}^{N}[p_t(F_k(\{\sigma\}))w_k(F_k(\{\sigma\}))-p_t(\{\sigma\})w_k(\{\sigma\})], \tag{23}$$

where we defined single spin flip operator F_k by

$$F_k(\{\sigma\})=(\sigma_1,\sigma_2,\ldots,-\sigma_k,\ldots,\sigma_N)=\{\sigma\}'. \tag{24}$$

Distribution $P_t(m,a)$, which is the probability that the system has macroscopic order parameters

$$m(\{\sigma\})\equiv\frac{1}{N}\sum_i \sigma_i,\quad a(\{\sigma\})\equiv\frac{1}{N}\sum_i \tau_i\sigma_i \tag{25}$$

at time t, is written in terms of the distribution $p_t(\{\sigma\})$ of the microscopic state $\{\sigma\}$ as

$$P_t(m,a)=\sum_\sigma p_t(\{\sigma\})\delta(m-m(\{\sigma\}))\delta(a-a(\{\sigma\})), \tag{26}$$

where $\delta(\cdots)$ is a delta function. Taking a derivative of $P_t(m,a)$ with respect to t and substituting Eq. (23) into this expression and making a Taylor expansion in powers of $2\sigma_k/N$ and $2\tau_k\sigma_k/N$ (the so-called *Kramers-Moyal expansion*), we obtain

$$\begin{aligned}\frac{dP_t(m,a)}{dt}=&\frac{\partial}{\partial m}P_t(m,a)\left\{m-\frac{\sum_\xi e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\times\int_{-\infty}^{\infty}Dx\tanh(Jm+h\tau x+h\tau_0\xi)\right\}\\&+\frac{\partial}{\partial a}P_t(m,a)\left\{a-\frac{\sum_\xi e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\times\int_{-\infty}^{\infty}Dx(\tau x+\tau_0\xi)\tanh(Jm+h\tau x+h\tau_0\xi)\right\}\\&+\mathcal{O}(N^{-1}).\end{aligned} \tag{27}$$

Thus, we derived the time-dependent distribution of macroscopic quantities from the microscopic master equation, Eq.

(23). Finally, we construct differential equations with respect to the macroscopic quantities m and a. Substituting a form of distribution

$$P_t(m,a)=\delta(m-m(t))\delta(a-a(t)) \tag{28}$$

into Eq. (27) and calculating some integrals, we obtain

$$\frac{dm}{dt}=-m+\frac{\sum_{\xi} e^{\beta_s m_0 \xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty} Dx \tanh(Jm+h\tau x+h\tau_0\xi), \tag{29}$$

$$\frac{da}{dt}=-a+\frac{\sum_{\xi} e^{\beta_s m_0 \xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty} Dx(\tau x+\tau_0\xi)\tanh(Jm+h\tau x + h\tau_0\xi). \tag{30}$$

These two equations describe a relaxation of the system to the equilibrium state. We should notice that the order parameter a is a slave variable in the sense that the order parameter m relaxes independently, whereas the relaxation of a depends on m. Therefore, the behavior of a is completely determined by m. For this reason, from now on, we disregard Eq. (30).

It is easy to see that in the limit of $t\to\infty$ and $dm/dt=0$, the saddle point equation (19) is recovered. As the overlap M is written in terms of m [see Eq. (20)], the time evolution of the overlap is obtained by substituting the time dependence of the magnetization $m(t)$ into the expression of M.

Using the same technique as the procedure to derive the differential equation with respect to m, the differential equation for the magnetization m_1 of the prior system $P(\{\sigma\})=\exp(J\Sigma_{ij}\sigma_i\sigma_j)/\Sigma_\sigma\exp(J\Sigma_{ij}\sigma_i\sigma_j)$ is obtained as

$$\frac{dm_1}{dt}=-m_1+\tanh(m_1 J). \tag{31}$$

Although in these equations we regard the hyperparameters J and h as constant variables, one should treat them as time-dependent parameters, that is, $J(t)$ and $h(t)$ from the viewpoint of hyperparameter estimation. Of course, details of the time dependence of $J(t)$ and $h(t)$ depend on a particular algorithm the of hyperparameter estimation. In the next section we investigate properties of hyperparameter estimation as a dynamical process of the coupling constant $J(t)$ and the field strength $h(t)$.

IV. HYPER-PARAMETER ESTIMATION

In Secs. II and III we investigated both static and dynamical properties of image restoration. From those results, we obtained hyperparameter dependence of the overlap explicitly. Moreover, for a particular constant hyperparameter set (J,h), we derived the differential equations which describe the relaxation of the system. As one of the authors reported in [4], if one fails to set the hyperparameters appropriately, the restored image gets worse than the degraded image. In practical situations, we do not know the optimal value of the hyperparameters before we carry out the MCMC simulations. Therefore, we need to determine the optimal value by using only information about the degraded image $\{\tau\}$. Of course, it is possible for us to construct some robust algorithms for hyperparameter tuning and several authors reported such algorithms based on *selective freezing* [13] or *quantum fluctuation* [14]. However, if one seeks the optimal restoration, hyperparameter estimation becomes a very important problem.

About ten years ago, Iba [12] studied the performance of the MML method with the assistance of the MCMC simulations for the same problem as ours. However, as he mentioned in his paper, the results are not enough to make its performance clear due to the difficulties of simulating the equilibrium state within reliable precision. With this fact in mind, in this section we calculate the marginal likelihood as a function of hyperparameters analytically. From the marginal likelihood, we derive Boltzmann machine-type learning equations and investigate their behavior quantitatively.

A. Maximum marginal likelihood method

In statistics, the maximum marginal likelihood (MML) method is used to infer hyperparameters appearing in the posterior distribution [1,7,15]. In the context of image restoration, marginal likelihood (the logarithm of marginal likelihood) is given by

$$\begin{aligned}-K(J,h:\{\xi,\tau\}) &\equiv \log\sum_{\sigma} P(\{\tau\}|\{\sigma\})P(\{\sigma\}) \\ &= \log\left(\sum_{\sigma} e^{J\Sigma_{ij}\sigma_i\sigma_j+h\Sigma_i\tau_i\sigma_i}\right)-\log Z_\Pi \\ &\quad -\log Z_L,\end{aligned} \tag{32}$$

where Z_Π and Z_L are given by Eq. (9). We should remember that Z_L is independent of $\{\sigma\}$ for both cases of the BSC and the GC. Usually, we attempt to maximize the marginal likelihood by using gradient descent with respect to J and h. This result leads to the following Boltzmann machine-type learning equations:

$$\begin{aligned}c_J\frac{dJ}{dt} &= -\frac{\partial K(J,h:\{\xi,\tau\})}{\partial J} \\ &= \frac{\sum_{\sigma}\left(\sum_{ij}\sigma_i\sigma_j\right)e^{J\Sigma_{ij}\sigma_i\sigma_j+h\Sigma_i\tau_i\sigma_i}}{\sum_{\sigma} e^{J\Sigma_{ij}\sigma_i\sigma_j+h\Sigma_i\tau_i\sigma_i}} \\ &\quad -\frac{\sum_{\sigma}\left(\sum_{ij}\sigma_i\sigma_j\right)e^{J\Sigma_{ij}\sigma_i\sigma_j}}{\sum_{\sigma} e^{J\Sigma_{ij}\sigma_i\sigma_j}},\end{aligned} \tag{33}$$

$$c_h \frac{dh}{dt} = -\frac{\partial K(J,h:\{\xi,\tau\})}{\partial h} = \frac{\sum_\sigma \left(\sum_i \tau_i\sigma_i\right) e^{J\Sigma_{ij}\sigma_i\sigma_j + h\Sigma_i\tau_i\sigma_i}}{\sum_\sigma e^{J\Sigma_{ij}\sigma_i\sigma_j + h\Sigma_i\tau_i\sigma_i}} - \frac{\partial \log Z_L}{\partial h}, \tag{34}$$

where c_J and c_h are relaxation times. Thus, by solving these two equations, we maximize the marginal likelihood $-K(J,h:\{\xi,\tau\})$ and obtain the values of hyperparameters as a fixed point of the equations. Then, we should notice that these two equations contain expectations of the quantities $\Sigma_{ij}\sigma_i\sigma_j$ and $\Sigma_i\tau_i\sigma_i$ over the posterior and the prior distributions. Therefore, when we solve Eqs. (33) and (34) numerically, we should calculate these expectations at each time step of the Euler method. Iba [12] carried out the MCMC method to calculate the expectations and evaluated time dependence of the hyperparameters J and h numerically. However, the accuracy of his computer simulation is not reliable because the time to simulate the equilibrium state is not enough. Accordingly, it is worthwhile to investigate the performance of the MML method analytically using the solvable model. In this section, we use the infinite range mean-field MRF model and solve the learning equations (33) and (34) exactly.

As our interest is an averaged performance of the MML method, we should calculate the averaged marginal likelihood,

$$-[K(J,h:\{\xi,\tau\})]_{\{\xi,\tau\}} = \left[\log\sum_\sigma e^{(J/N)\Sigma_{ij}\sigma_i\sigma_j + h\Sigma_i\tau_i\sigma_i}\right]_{\{\xi,\tau\}} - \left[\log\sum_\sigma e^{(J/N)\Sigma_{ij}\sigma_i\sigma_j}\right]_{\{\xi,\tau\}} - [\log Z_L]_{\{\xi,\tau\}}, \tag{35}$$

where the bracket $[\cdots]_{\{\xi,\tau\}}$ means the average over the distribution $P(\{\tau\}|\{\xi\})P(\{\xi\})$ and the sum $\Sigma_{ij}(\cdots)$ should be carried out for all pairs of pixels. We should keep in mind that we rescaled the coupling constant as J/N to make the averaged marginal likelihood (difference of free energy) of order N. In general, it is hard to carry out this kind of average, namely, $[\log Z]_{\{\xi,\tau\}}$. Then, we replace the average with an average of the nth moment of Z, that is, Z^n by using

$$[\log Z]_{\{\xi,\tau\}} = \lim_{n\to 0}\frac{[Z^n]_{\{\xi,\tau\}} - 1}{n}. \tag{36}$$

This is referred to as the *replica method* [11]. By using the replica method, we obtain the averaged marginal likelihood per pixel as

$$-\frac{[K(J,h:\{\xi,\tau\})]_{\{\xi,\tau\}}}{N} = -\frac{J}{2}m^2 + \frac{\sum_\xi e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty} Dx \log[2\cosh(Jm + h\tau x + h\tau_0\xi)] + \frac{J}{2}m_1^2 - \log 2\cosh(m_1 J) + \frac{\tau_0}{2\tau^2} - \frac{\tau^2 h^2}{2} \equiv -K(J,h), \tag{37}$$

where m and m_1 are magnetizations of the spin systems described by the posterior and the prior, respectively. It should be noticed that as we used the GC [Eqs. (3) and (4)], the average $[\log Z_L]_{\{\xi,\tau\}}$ simply led to $(Nh^2/2) - (N\tau_0/2\tau^2)$ [see Eq. (11)].

In Fig. 2, we plot the averaged marginal likelihood as a function of J and h. In this figure we see that the averaged marginal likelihood takes its maximum when we choose the hyperparameters (J,h) so as to be identical to their true values $(\beta_s = 1/T_s = 1.1, \beta_\tau = \tau_0/\tau^2 = 1)$ (we set $\tau_0 = \tau = 1, T_s = 0.9$). This fact is easily checked by the following inequality [16]:

$$\begin{aligned}
&\{-[K(\beta_s,\beta_\tau:\{\xi,\tau\})]_{\{\xi,\tau\}}\} - \{-[K(J,h:\{\xi,\tau\})]_{\{\xi,\tau\}}\} \\
&= \sum_{\xi,\tau} P_{\beta_\tau}(\{\tau\}|\{\xi\}) P_{\beta_s}(\{\xi\}) \log \sum_\sigma P_{\beta_\tau}(\{\tau\}|\{\sigma\}) \\
&\quad \times P_{\beta_s}(\{\sigma\}) - \sum_{\xi,\tau} P_{\beta_\tau}(\{\tau\}|\{\xi\}) \\
&\quad \times P_{\beta_s}(\{\xi\}) \log \sum_\sigma P_h(\{\tau\}|\{\sigma\}) P_J(\{\sigma\}) \\
&= \sum_\tau P_{\beta_s,\beta_\tau}(\{\tau\}) \log(P_{\beta_s,\beta_\tau}(\{\tau\})/P_{J,h}(\{\tau\})) \geqslant 0,
\end{aligned} \tag{38}$$

where we used the non-negativity of *Kullback-Libeler information* and we defined

$$\begin{aligned}
P_X(\{\tau\}|\{\sigma\}) &\equiv \frac{\exp\left(X\sum_i \tau_i\sigma_i\right)}{\sum_\tau \exp\left(X\sum_i \tau_i\sigma_i\right)}, \\
P_X(\{\tau\}|\{\xi\}) &\equiv \frac{\exp\left(X\sum_i \tau_i\xi_i\right)}{\sum_\tau \exp\left(X\sum_i \tau_i\xi_i\right)}, \\
P_Y(\{\sigma\}) &\equiv \frac{\exp\left(Y\sum_{ij} \sigma_i\sigma_j\right)}{\sum_\sigma \exp\left(Y\sum_{ij} \sigma_i\sigma_j\right)},
\end{aligned} \tag{39}$$

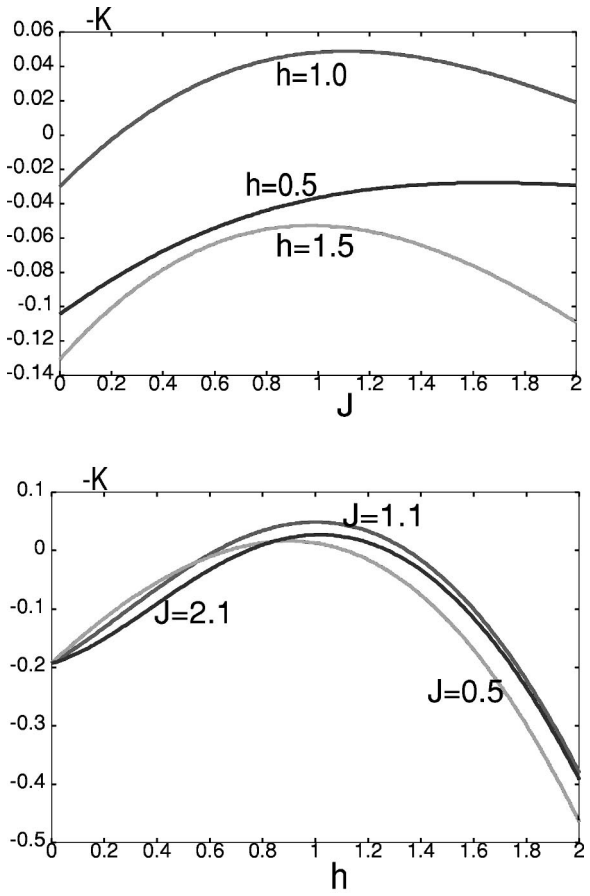

FIG. 2. J dependence of the averaged marginal likelihood $-K$ (upper figure). We set $h=0.5,1$ and $h=1.5$. We see that $-K$ takes its maximum when we choose J,h as $J=1.1(=1/T_s)$ and $h=\beta_\tau=1$. h dependence of the averaged marginal likelihood $-K$ (lower figure). We set $J=0.5,1$ and $J=2.1$. We see that $-K$ takes its maximum when we choose J,h as $J=1.1=(1/T_s)$ and $h=\beta_\tau=1$. For both figures, we chose (m,m_1) as a solution of Eq. (19) and $m_1=\tanh(Jm_1)$ for $J=1/T_s$ and $h=\beta_\tau$.

$$P_Y(\{\xi\}) \equiv \frac{\exp\left(Y\sum_{ij}\xi_i\xi_j\right)}{\sum_{\xi}\exp\left(Y\sum_{ij}\xi_i\xi_j\right)}, \tag{40}$$

$$P_{X,Y}(\{\tau\}) \equiv \sum_{\sigma} P_X(\{\tau\}|\{\sigma\})P_Y(\{\sigma\})$$

$$= \sum_{\xi} P_X(\{\tau\}|\{\xi\})P_Y(\{\xi\}). \tag{41}$$

Thus, we confirm that our mean-field model is not against this general inequality. We should mention that the static properties of the hyperparameter estimation were investigated by several authors using the generalized Gaussian model [17], mean-field approximation [1], and cluster variation method [18].

For the marginal likelihood (35), averaged learning equations with respect to J and h are obtained by the gradient descent

$$c_J\frac{dJ}{dt} = -\left[\frac{\partial K(J,h:\{\xi,\tau\})}{\partial J}\right]_{\{\xi,\tau\}},$$

$$c_h\frac{\partial h}{\partial t} = -\left[\frac{\partial K(J,h:\{\xi,\tau\})}{\partial h}\right]_{\{\xi,\tau\}}. \tag{42}$$

The right-hand sides of the above equations are also evaluated by the replica method. After some algebra, we obtain

$$c_J\frac{dJ}{dt} = -\frac{m^2}{2} + m\frac{\sum_{\xi} e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty}Dx \times\tanh(Jm+h\tau x+h\tau_0\xi) + \frac{m_1^2}{2} - m_1\tanh(m_1 J), \tag{43}$$

$$c_h\frac{dh}{dt} = \frac{\sum_{\xi} e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty}Dx(\tau x+\tau_0\xi) \times\tanh(Jm+h\tau x+h\tau_0\xi) - \tau^2 h, \tag{44}$$

where we should remember that m and m_1 obey the differential equations

$$\frac{dm}{dt} = -m + \frac{\sum_{\xi} e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty}Dx\tanh(Jm+h\tau x+h\tau_0\xi), \tag{45}$$

$$\frac{dm_1}{dt} = -m_1 + \tanh(m_1 J). \tag{46}$$

By solving these coupled equations, we obtain time dependences of the hyperparameters $J(t),h(t)$ and relaxation process of the systems, namely, $m(t),m_1(t)$. In this paper we fix the relaxation times as $c_J=c_h=1$.

In Fig. 3 we plot time dependences of the hyperparameters J,h and order parameters m,m_1. From this figure we see that the final state of the hyperparameters is optimal, namely, $(J_*,h_*)\equiv(1/T_s,\beta_\tau=\tau_0/\tau^2)=(1.1,1)$ and this convergent point is independent of the initial conditions. Time evolutions of the overlap M are also plotted in Fig. 4 (upper figure). We find that the overlap M converges to the best possible value in Fig. 1. In Fig. 5 we plot flows of hyperparameter J-h. From this figure, we find that each flow does not take the shortest path to the solution and goes a long way around the solution.

B. EM algorithm

In the preceding section we investigated the process of the MML method by gradient decent as a dynamics. In this sec-

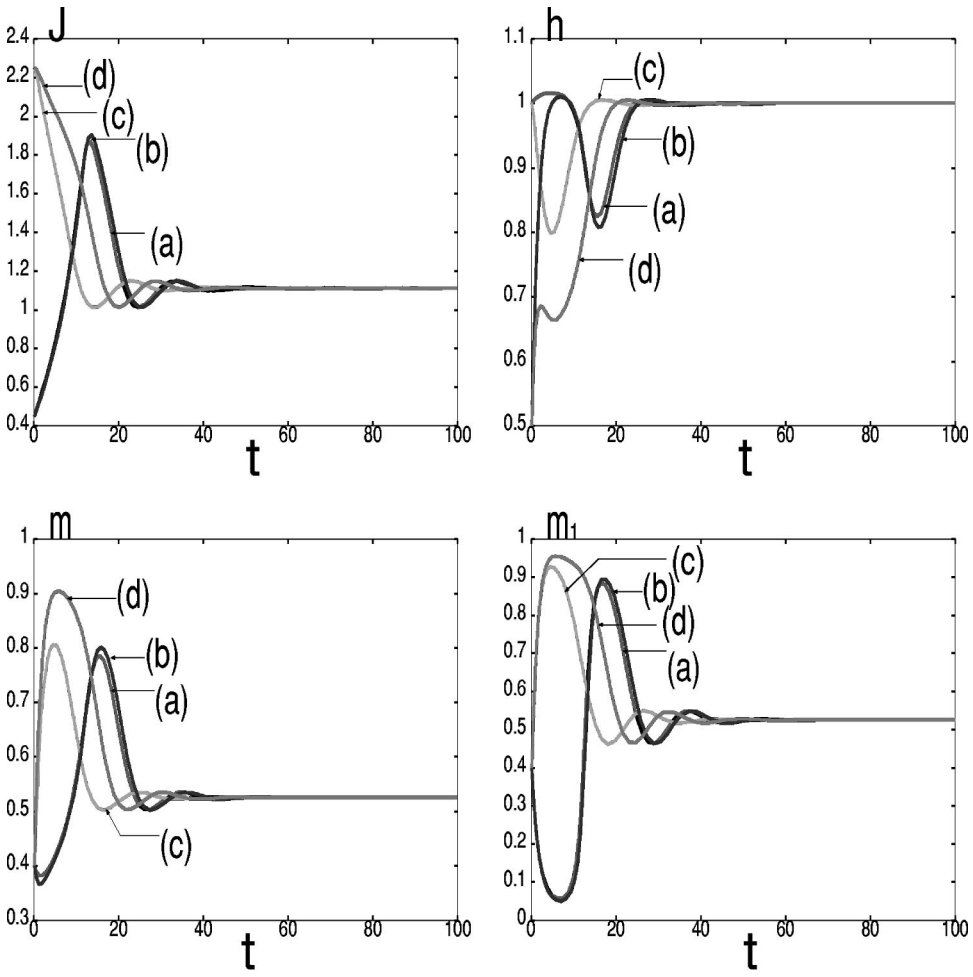

FIG. 3. From the upper left to the lower right, time dependences of the hyperparameters J, h and the magnetizations m, m_1 are plotted. In each graph, we choose the initial condition (a) $J(0)=0.45, h(0)=1, m(0)=m_1(0)=0.4$; (b) $J(0)=0.45, h(0)=0.5, m(0)=m_1(0)=0.4$; (c) $J(0)=2.25, h(0)=1, m(0)=m_1(0)=0.4$; (d) $J(0)=2.25, h(0)=0.5, m(0)=m_1(0)=0.4$. We set true values of the hyperparameters $T_s=0.9$, $\beta_\tau=1$.

tion we analyze the performance of the *EM algorithm* [8] as another candidate to maximize the marginal likelihood.

In the EM algorithm, we first average the logarithmic-likelihood function

$$\log P(\{\tau\}|\{\sigma\})P(\{\sigma\}) \equiv \frac{J}{N}\sum_{ij}\sigma_i\sigma_j + h\sum_i \tau_i\sigma_i - \log\sum_\sigma \exp\left(\frac{J}{N}\sum_{ij}\sigma_i\sigma_j\right) + \frac{N\tau_0}{2\tau^2} - \frac{N\tau^2h^2}{2} \tag{47}$$

over the time-dependent posterior distribution

$$P_t(\{\sigma\}|\{\tau\}) \equiv \frac{e^{(J_t/N)\Sigma_{ij}\sigma_i\sigma_j + h_t\Sigma_i\tau_i\sigma_i}}{\sum_\sigma e^{(J_t/N)\Sigma_{ij}\sigma_i\sigma_j + h_t\Sigma_i\tau_i\sigma_i}}. \tag{48}$$

This average is referred to as a Q function. As we are interested in the averaged behavior of the Q function, we need the following averaged Q function:

$$\begin{aligned} Q(J,h|J_t,h_t) &\equiv \left[\sum_\sigma P_t(\{\sigma\}|\{\tau\})\log P(\{\tau\}|\{\sigma\})P(\{\sigma\})\right]_{\{\xi,\tau\}} \\ &= J\left[\frac{\sum_\sigma\left(\sum_{ij}\sigma_i\sigma_j\right)e^{(J_t/N)\Sigma_{ij}\sigma_i\sigma_j + h_t\Sigma_i\tau_i\sigma_i}}{\sum_\sigma e^{(J_t/N)\Sigma_{ij}\sigma_i\sigma_j + h_t\Sigma_i\tau_i\sigma_i}}\right]_{\{\xi,\tau\}} \\ &\quad + h\left[\frac{\sum_\sigma\left(\sum_i\tau_i\sigma_i\right)e^{(J_t/N)\Sigma_{ij}\sigma_i\sigma_j + h_t\Sigma_i\tau_i\sigma_i}}{\sum_\sigma e^{(J_t/N)\Sigma_{ij}\sigma_i\sigma_j + h_t\Sigma_i\tau_i\sigma_i}}\right]_{\{\xi,\tau\}} \\ &\quad - \log\sum_\sigma\exp\left(\frac{J}{N}\sum_{ij}\sigma_i\sigma_j\right) + \frac{N\tau_0^2}{2\tau^2} - \frac{N\tau^2h^2}{2}, \end{aligned} \tag{49}$$

where we divided the coupling constants J and J_t by N to take a proper thermodynamic limit. Then, the EM algorithm is summarized as follows.

(i) *Step 1*. Set initial values of the hyperparameters J_0, h_0, and $t \leftarrow 0$.

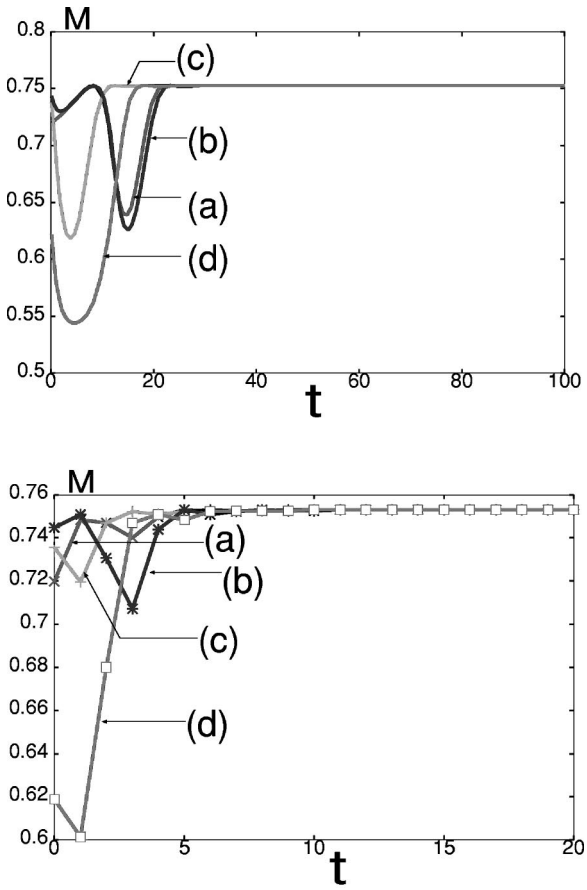

FIG. 4. Time dependences of the overlap M for the case of the MML by gradient descent (upper figure) and the case of the EM algorithm (lower figure). For both cases, we choose the initial condition as (a) $J(0)=0.45$, $h(0)=1$, $m(0)=m_1(0)=0.4$; (b) $J(0)=0.45$, $h(0)=0.5$, $m(0)=m_1(0)=0.4$; (c) $J(0)=2.25$, $h(0)=1$, $m(0)=m_1(0)=0.4$; (d) $J(0)=2.25$, $h(0)=0.5$, $m(0)=m_1(0)=0.4$. We set true values of the hyperparameters $T_s=0.9$, $\beta_\tau=1$. We see that for both cases, the optimal overlap $M_{\rm opt}$ is obtained as a fixed point of the dynamics.

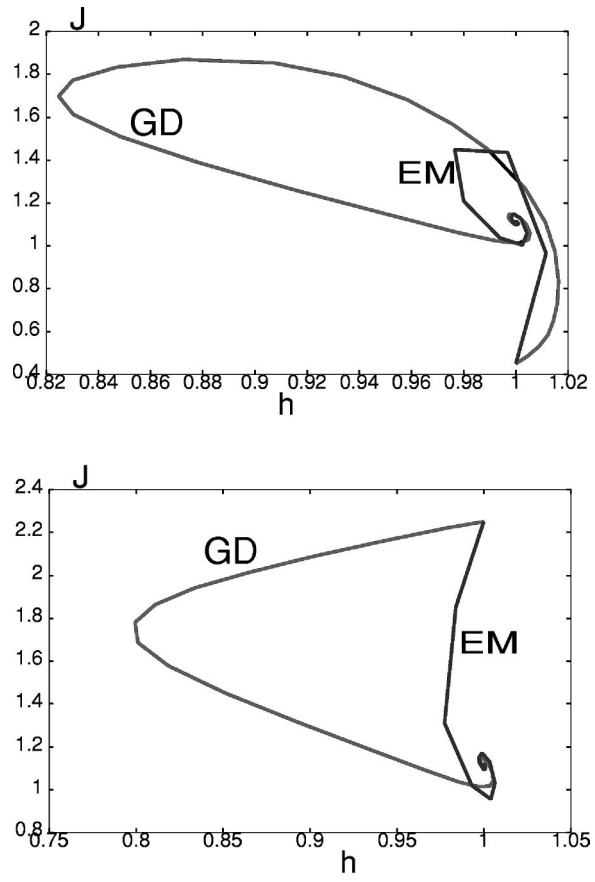

FIG. 5. Flows in the hyperparameter space (J,h). We set the initial conditions $J(0)=J_0=0.45$, $h(0)=h_0=1$, and $m(0)=m_1(0)=0.4$ (upper figure) and $J(0)=J_0=2.25$, $h(0)=h_0=1$ and $m(0)=m_1(0)=0.4$ (lower figure). True values of the hyperparameters are $J_*=1/T_s=1.1$, $h_*=\beta_\tau=1$. For the case of gradient descent (GD), the flows go a long way around the solution $(J_*,h_*)=(1.1,1)$. In order to compare the MML by gradient descent with the EM algorithm, we also plot flows of the EM algorithm (EM). We see that the EM algorithm takes shorter paths than the MML by gradient descent.

(ii) *Step 2*. Iterate the following E (expectation) and M (maximization) steps until an appropriate convergence condition is satisfied. For the E step: calculate $Q(J,h|J_t,h_t)$. For the M step: update J_t and h_t by

$$J_{t+1}=\mathrm{argmax}_J Q(J,h|J_t,h_t)$$

$$h_{t+1}=\mathrm{argmax}_h Q(J,h|J_t,h_t),$$

and

$$t\leftarrow t+1.$$

For our infinite range mean-field MRF model, the averages $[\cdots]_{\{\xi,\tau\}}$ in Eq. (49) are calculated by using the replica method and we obtain

$$\frac{Q(J,h|J_t,h_t)}{N}=-\frac{Jm(t)^2}{2}+\frac{Jm(t)\sum_\xi e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\times\int_{-\infty}^{\infty}Dx\tanh[J_t m(t)+h_t\tau x+h_t\tau_0\xi]+\frac{h\sum_\xi e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty}Dx(\tau x+\tau_0\xi)\times\tanh[J_t m(t)+h_t\tau x+h_t\tau_0\xi]+\frac{J}{2}m_1(t)^2-\log 2\cosh[m_1(t)J]+\frac{\tau_0^2}{2\tau^2}-\frac{\tau^2h^2}{2}. \quad (50)$$

At the next time step, J_{t+1} and h_{t+1} are given by the condi-

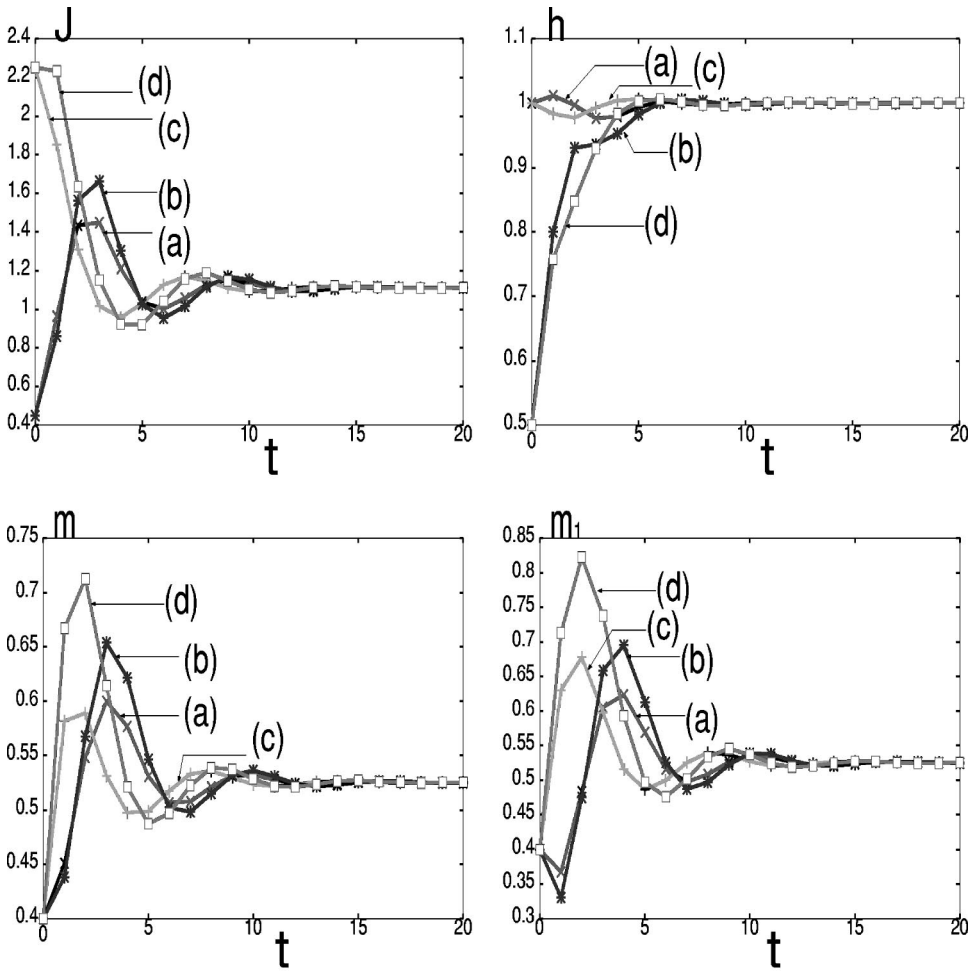

FIG. 6. From the upper left to the lower right, time dependences of the hyperparameters J, h and the magnetizations m, m_1 for the EM algorithm are plotted. In each graph, we choose the initial condition (a) $J_0=0.45$, $h_0=1$, $m(0)=m_1(0)=0.4$; (b) $J_0=0.45$, $h_0=0.5$, $m(0)=m_1(0)=0.4$; (c) $J_0=2.25$, $h_0=1$, $m(0)=m_1(0)=0.4$; (d) $J_0=2.25$, $h_0=0.5$, $m(0)=m_1(0)=0.4$. We set true values of the hyperparameters $T_s=0.9$, $\beta_\tau=1$.

tions $\partial Q/\partial J=0$ and $\partial Q/\partial h=0$. These two conditions lead to nonlinear maps

$$J_{t+1}=\frac{1}{m(t)}\tanh^{-1}\left[-\frac{\{m(t)^2-m_1(t)^2\}^2}{2m_1(t)}+\frac{m(t)\sum_{\xi}e^{\beta_s m_0\xi}}{2m_1(t)\cosh(\beta_s m_0)}\times\int_{-\infty}^{\infty}Dx\tanh[J_t m(t)+h_t\tau x+h_t\tau\xi]\right],\quad(51)$$

$$h_{t+1}=\frac{\sum_{\xi}e^{\beta_s m_0\xi}}{2\tau^2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty}Dx(\tau x+\tau_0\xi)\times\tanh[J_t m(t)+h_t\tau x+h_t\tau_0\xi].\quad(52)$$

In the above nonlinear maps, $m(t)$ and $m_1(t)$ are time-dependent magnetizations for the systems described by the posterior $P(\{\sigma\}|\{\tau\})$ and the the prior $P(\{\sigma\})$, respectively. By using mean-field treatment, we obtain nonlinear maps with respect to $m(t)$ and $m_1(t)$ as

$$m(t+1)=\frac{\sum_{\xi}e^{\beta_s m_0\xi}}{2\cosh(\beta_s m_0)}\int_{-\infty}^{\infty}Dx\times\tanh[J_t m(t)+h_t\tau x+h_t\tau_0\xi],\quad(53)$$

$$m_1(t+1)=\tanh[J_t m_1(t)].\quad(54)$$

By solving these nonlinear maps, Eqs. (51)–(54), we obtain the time dependence of the hyperparameters J_t,h_t and the magnetizations $m(t)$,$m_1(t)$. We plot the results in Fig. 4 (lower figure), Fig. 5, and Fig. 6. From these figures we see that both the MML method by gradient descent and the EM algorithm obtain the optimal hyperparameters $(J_*,h_*)=(1.1,1)$; however, the EM algorithm shows faster convergence than the MML by gradient descent. In addition, the flows of the EM algorithm in the hyperparameter space are shorter than those of the MML by gradient descent. From the posterior distribution appearing in the Q function (49), we see that performance of the EM algorithm highly depends on the initial choice of the hyperparameters J_0 and h_0. Therefore, for the systems which have lots of local minima, the final solution is sensitive to the initial condition on the hy-

perparameters. However, for our model system (the infinite range random field Ising model), there is no local minima in the marginal likelihood function. As a result, the final state of the EM algorithm is independent of the initial conditions.

V. SUMMARY

In this paper we investigated dynamical properties of image restoration by using statistical mechanics. We introduced an infinite range mean-field version of the MRF model and solved it analytically. We derived differential equations with respect to the macroscopic order parameters from the microscopic Master equation. We also studied dynamics of hyperparameter estimation in the context of the maximum marginal likelihood method by using gradient descent and the EM algorithm. For the MML method by gradient descent, Boltzmann machine-type learning equations were evaluated analytically by the replica method. On the other hand, the EM algorithm led to nonlinear maps and these maps were also evaluated analytically. We compared these two algorithms and found that for both algorithms we obtain the optimal hyperparameters. We also found that the speed of convergence for the EM algorithm is faster than that of the MML method by gradient descent. In addition, the paths to the solution in hyperparameter space by the EM algorithm are shorter than those of the MML by gradient descent. Thus, in this paper, we could compare two different methods to estimate hyperparameters without any computer simulations. Our analytical treatments are applicable to studies of performance for the other method, including the *deterministic annealing EM algorithm* [19,20]. Moreover, besides image restoration, our approach is useful for the other problems, for example, learning by Bayesian neural networks [21,22], time series predictions [23], or the density estimation problem [24].

ACKNOWLEDGMENTS

We thank Hidetoshi Nishimori, Masato Okada, Yukito Iba, and David Saad for fruitful discussions. Our special thanks are due to Toshiyuki Tanaka for useful discussions and comments.

[1] J.M. Pryce and A.D. Bruce, J. Phys. A **28**, 511 (1995).
[2] H. Nishimori and K.Y.M. Wong, Phys. Rev. E **60**, 132 (1999).
[3] D.M. Carlucci and J. Inoue, Phys. Rev. E **60**, 2547 (1999).
[4] J. Inoue and D.M. Carlucci, Phys. Rev. E **64**, 036121 (2001).
[5] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).
[6] J. Marroquin, S. Mitter, and T. Poggio, J. Am. Stat. Assoc. **82**, 76 (1989).
[7] S. Geman and D.E. McClure, Bull. Int. Statist. Inst. **52**, 5 (1987).
[8] A.P. Dempster, N.M. Laird, and D.B. Rubin, J. R. Statist., Ser. B (methodological) **39**, 1 (1977).
[9] S. Kirkpatrick, G.D. Gellatt, Jr, and M.P. Vecchi, Science **220**, 671 (1983).
[10] D. Geiger and F. Girosi, IEEE Trans. Pattern Anal. Mach. Intell. **13**, 401 (1991).
[11] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[12] Y. Iba, Proc. Inst. Statist. Math. (in Japanese) **39**, 1 (1991).
[13] K.Y.M. Wong and H. Nishimori, Phys. Rev. E **62**, 179 (2000).
[14] J. Inoue, Phys. Rev. E **63**, 046114 (2001).
[15] Z. Zhou, R.M. Leahy, and J. Qi, IEEE Trans. Image Process. **6**, 844 (1997).
[16] Y. Iba, J. Phys. A **32**, 3875 (1999).
[17] K. Tanaka and J. Inoue, Technical report of the Institute of Electronics, Information and Communication Engineers (in Japanese), Report No. PRMU2000-125 (2000), p. 41.
[18] K. Tanaka, Trans. Jpn. Soc. for Artificial Intell. **16**, 246 (2001).
[19] R.L. Streit and T.E. Luginbuhl, IEEE Trans. Neural Netw. **5**, 764 (1994).
[20] N. Ueda and R. Nakano, Neural Networks **11**, 271 (1998).
[21] D. J. C. Mackay, in *Models of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer-Verlag, Berlin, 1996).
[22] B. J. Frey, *Graphical Model for Machine Learning and Digital Communication* (MIT, Combridge, MA, 1998).
[23] T. Matsumoto, Y. Nakajima, M. Saito, J. Sugi and H. Hamagishi, IEEE Trans. Signal Processing (to be published).
[24] N. Barkai and H. Sompolinsky, Phys. Rev. E **50**, 1766 (1994).