Mean field theory of EM algorithm for Bayesian grey scale image restoration

2003 J. Phys. A: Math. Gen. 36 10997

(http://iopscience.iop.org/0305-4470/36/43/023)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 10997–11010 PII: S0305-4470(03)60852-8

Mean field theory of EM algorithm for Bayesian grey scale image restoration

Jun-ichi Inoue[1] **and Kazuyuki Tanaka**[2]

[1] Complex Systems Engineering, Graduate School of Engineering, Hokkaido University, N13-W8, Kita-ku, Sapporo 060-8028, Japan
[2] Department of Computer and Mathematical Science, Graduate School of Information Sciences, Tohoku University, Aramaki-aza-aoba 04, Aoba-ku Sendai 980-8579, Japan

E-mail: j_inoue@complex.eng.hokudai.ac.jp

Received 14 March 2003, in final form 23 May 2003
Published 15 October 2003
Online at stacks.iop.org/JPhysA/36/10997

Abstract

The EM algorithm for the Bayesian grey scale image restoration is investigated in the framework of the mean field theory. Our model system is identical to the infinite range random field Q-Ising model. The maximum marginal likelihood method is applied to the determination of hyper-parameters. We calculate both the data-averaged mean square error between the original image and its maximizer of posterior marginal estimate, and the data-averaged marginal likelihood function exactly. After evaluating the hyper-parameter dependence of the data-averaged marginal likelihood function, we derive the EM algorithm which updates the hyper-parameters to obtain the maximum likelihood estimate analytically. The time evolutions of the hyper-parameters and so-called Q function are obtained. The relation between the speed of convergence of the hyper-parameters and the shape of the Q function is explained from the viewpoint of dynamics.

PACS numbers: 02.50−r, 05.20−y, 05.50+q

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Recently, a statistical-mechanical approach to the problems of information processing was investigated from the viewpoint of Bayesian inference [1]. As a typical example of the Bayesian inferences from an incomplete data set, image restoration has been investigated by engineers, applied mathematicians and statistical physicists [2, 3]. Among these studies, analysis of the infinite range Markov random field model has played an important role to grasp some essential features of the model systems. By using this artificial model, Nishimori and Wong [4] obtained hyper-parameter dependence of the overlap between original and restored

images explicitly in the context of the maximizer of posterior marginal (MPM) estimation for black and white images. For grey scale image restoration, which is more important from a practical point of view, Carlucci and Inoue [6] and Inoue and Carlucci [7] tried to formulate the problem by the chiral Potts spin glass model and the $\mathcal{Q}$-Ising spin glass model [8], respectively.

In the Bayesian inference context, we should detect the optimal hyper-parameters from observable data (the degraded image in the context of image restoration) and this is one of the essential problems concerning image restoration. For this problem, the maximum marginal likelihood method is applied to decide the maximum likelihood estimate of the hyper-parameters. For black and white image restoration, the present authors calculated the data-averaged marginal likelihood function explicitly by means of the mean-field Ising model and investigated the hyper-parameter dependence of the likelihood function explicitly [9]. They also constructed the gradient descent and the EM algorithm [10] to maximize the marginal likelihood function and compared these two methods.

In this paper, we extend our work [9] to a more realistic case, namely, the problem of grey scale image restoration by using the $\mathcal{Q}$-Ising model [6]. We focus on the dynamics of the EM algorithm and investigate the time evolution of the hyper-parameters, Q function analytically. Slow convergence by the EM algorithm along a specific hyper-parameter direction is explained from the viewpoint of time evolution of the Q function.

This paper is organized as follows. In section 2, we define the model system and explain the Bayesian grey scale image restoration using the $\mathcal{Q}$-Ising model. In section 3, the marginal likelihood function is introduced and we explain the criterion of the maximum marginal likelihood function to infer the hyper-parameters. In section 4, we calculate the averaged marginal likelihood function to investigate a typical performance of the maximum marginal likelihood method. The hyper-parameter dependence of the averaged marginal likelihood function is derived for the cases $\mathcal{Q} = 3$ and 4. In section 5, the EM algorithm is applied to the problem. For the mean-field model, the update rules of the hyper-parameters are obtained analytically. Flows of the hyper-parameters, the time dependences of the hyper-parameters, the Q function and the averaged marginal likelihood are given explicitly. The final section is devoted to a summary.

2. Definitions of the mean-field $\mathcal{Q}$-Ising system

In this section, we explain the definitions of our image restoration system. The original images $\{\xi\} \equiv (\xi_1, \xi_2, \ldots, \xi_N)$ are given as snapshots from the Gibbs distribution of the ferromagnetic $\mathcal{Q}$-Ising model [7]:

$$P_{\beta_s}(\{\xi\}) = \frac{\exp\left[-(\beta_s/2N)\sum_{ij}(\xi_i - \xi_j)^2\right]}{Z_0(\beta_s)} \tag{1}$$

where each pixel ξ_i takes $\mathcal{Q}$ values, namely, $\xi_i = 0, 1, 2, \ldots, \mathcal{Q} - 1$. The partition function $Z_0(\beta_s)$ is a normalization constant of the distribution (1):

$$Z_0(\beta_s) \equiv \mathrm{tr}_{\{\xi\}} \exp\left[-(\beta_s/2N)\sum_{ij}(\xi_i - \xi_j)^2\right] \tag{2}$$

where $\mathrm{tr}_{\{\xi\}}(\cdots)$ means $\sum_{\xi_1=0}^{\mathcal{Q}-1}\sum_{\xi_2=0}^{\mathcal{Q}-1}\cdots\sum_{\xi_N=0}^{\mathcal{Q}-1}(\cdots)$. We should keep in mind that the energy appearing in the shoulder of the exponential of equation (1) is divided by N because we consider the infinite range model in which every pixel is connected.

For this original image, a degrading process, namely, the process from an original image $\{\xi\}$ to a degraded image $\{\tau\} \equiv (\tau_1, \tau_2, \ldots, \tau_N)$ is described by the following conditional probability (what we call the *Gaussian channel*):

$$P_{a_0,a_\tau}(\{\tau\}|\{\xi\}) = \frac{\exp\left[-\left(1/2a_\tau^2\right)\sum_i(\tau_i - a_0\xi_i)^2\right]}{\sqrt{2\pi}a_\tau}. \tag{3}$$

Thus, the original image $\{\xi\}$ is degraded by the Gaussian channel with mean $\{a_0\xi\} \equiv (a_0\xi_1, a_0\xi_2, \ldots, a_0\xi_N)$ and variance a_τ^2.

In order to restore the original image $\{\xi\}$ from the observable degraded image $\{\tau\}$, we construct the posterior distribution by means of the Bayes rule:

$$\begin{aligned} P_{\beta,h}(\{\sigma\}|\{\tau\}) &= \frac{P_h(\{\tau\}|\{\sigma\})P_\beta(\{\sigma\})}{\mathrm{tr}_{\{\sigma\}}P_h(\{\sigma\}|\{\tau\})P_\beta(\{\sigma\})} \\ &= \frac{\exp\left[-h\sum_i(\tau_i - \sigma_i)^2 - (\beta/2N)\sum_{ij}(\sigma_i - \sigma_j)^2\right]}{\mathrm{tr}_{\{\sigma\}}\exp\left[-h\sum_i(\tau_i - \sigma_i)^2 - (\beta/2N)\sum_{ij}(\sigma_i - \sigma_j)^2\right]} \end{aligned} \tag{4}$$

where $\{\sigma\} \equiv (\sigma_1, \sigma_2, \ldots, \sigma_N)$ means estimates of the original images $\{\xi\}$. $P_h(\{\tau\}|\{\sigma\})$ and $P_\beta(\{\sigma\})$ are regarded as a model of the channel (3) and a model of the distribution (1) (so-called *prior distribution*), respectively. The hyper-parameters h and β specify these model distributions and we should infer the values from incomplete data sets, namely, the degraded images $\{\tau\}$. We should bear in mind that it is possible for us to send more information about the original image by using the *parity check* $\xi_i\xi_j(\forall_{i\neq j})$ besides the sequence of the original image $\{\xi\}$. Then, the system is described by the $\mathcal{Q}$-Ising spin glass model [7, 8]. However, in this paper, we restrict ourselves to the case without any extra information, and the system is identical to the random field $\mathcal{Q}$-Ising model.

For this posterior distribution (4), the MPM estimate of the ith pixel $\hat{\xi}_i$ is given by

$$\hat{\xi}_i = \Omega(\langle\sigma_i\rangle) = \sum_{k=0}^{Q-1}\left[\Theta\left(\langle\sigma_i\rangle - \frac{2k-1}{2}\right) - \Theta\left(\langle\sigma_i\rangle - \frac{2k+1}{2}\right)\right] \tag{5}$$

$$\langle\sigma_i\rangle = \frac{\mathrm{tr}_{\{\sigma\}}\sigma_i\exp\left[-h\sum_i(\tau_i - \sigma_i)^2 - (\beta/2N)\sum_{ij}(\sigma_i - \sigma_j)^2\right]}{\mathrm{tr}_{\{\sigma\}}\exp\left[-h\sum_i(\tau_i - \sigma_i)^2 - (\beta/2N)\sum_{ij}(\sigma_i - \sigma_j)^2\right]} \tag{6}$$

where Θ is a step function defined by

$$\Theta(x) = \begin{cases} 1 & (x \geqslant 0) \\ 0 & (x < 0). \end{cases} \tag{7}$$

The reader should keep in mind that the estimate is regarded as $\hat{\xi}_i = 0$ if $\langle\sigma_i\rangle$ is smaller than $-1/2$ and $\hat{\xi}_i = \mathcal{Q}$ if $\langle\sigma_i\rangle$ is greater than $\mathcal{Q}/2$.

The quality of the restoration is measured by the following mean square error:

$$D = \frac{1}{2N}\sum_i(\xi_i - \hat{\xi}_i)^2. \tag{8}$$

For this image restoration system represented by the $\mathcal{Q}$-Ising model, Inoue and Carlucci [7] investigated the averaged performance of the MPM estimation by using the replica method and found that the mean square error takes its minimum at the true values of the hyper-parameters. However, in general, true values or optimal values of the hyper-parameter cannot be obtained before we calculate the estimate $\{\hat{\xi}\} \equiv (\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_N)$. For this reason, estimation of the hyper-parameters is needed in order to achieve the best possible Bayesian image restoration.

In the next section, we explain how we estimate the hyper-parameters in the context of the maximum marginal likelihood criterion and investigate its averaged case performance with the assistance of a statistical-mechanical technique.

3. Marginal likelihood function

As we explained in the previous section, hyper-parameter estimation is essential in the context of a Bayesian image restoration. When one attempts to infer the hyper-parameters, some appropriate criteria are needed. For such a criterion, one may use the minimum mean square error criterion. However, unfortunately, we cannot use it in practice. This is because, as we saw in equation (8), the mean square error contains the original image and we need it when we evaluate the minimum of the mean square error.

In statistics, the so-called marginal likelihood function is used as a cost function to attempt to maximize for decision of the hyper-parameters. In this section, we evaluate the maximum marginal likelihood criterion for the hyper-parameter estimation in our model system.

The marginal likelihood we use here is defined as follows:

$$\begin{aligned} -K(\beta, h : \{\xi, \tau\}) &\equiv \log \mathrm{tr}_{\{\sigma\}} P_h(\{\tau\}|\{\sigma\}) P_\beta(\{\sigma\}) \\ &= \log \mathrm{tr}_{\{\sigma\}} \frac{\exp\left[-h \sum_i (\sigma_i - \tau_i)^2 - (\beta/2N) \sum_{ij} (\sigma_i - \sigma_j)^2\right]}{Z_\Pi(\beta) Z_L(h)} \end{aligned} \tag{9}$$

where the following two partition functions are defined:

$$Z_L(h) \equiv \mathrm{tr}_{\{\tau\}} \exp\left[-h \sum_i (\sigma_i - \tau_i)^2\right] = \left(\frac{\pi}{h}\right)^{\frac{N}{2}} \tag{10}$$

$$Z_\Pi(\beta) = \mathrm{tr}_{\{\sigma\}} \exp\left[-(\beta/2N) \sum_{ij} (\sigma_i - \sigma_j)^2\right] = \exp\left[N\left(-\beta m_1^2 + \log \mathrm{tr}_\sigma \, \mathrm{e}^{2\beta m_1 \sigma - \beta\sigma^2}\right)\right] \tag{11}$$

where $\mathrm{tr}_\sigma F(\sigma)$ means $\sum_{\sigma=0}^{\mathcal{Q}-1} F(\sigma)$ for arbitrary function F. Magnetization m_1 obeys the saddle point equation

$$m_1 = \frac{\mathrm{tr}_\sigma \sigma \exp(2\beta m_1 \sigma - \beta\sigma^2)}{\mathrm{tr}_\sigma \exp(2\beta m_1 \sigma - \beta\sigma^2)}. \tag{12}$$

Thus, the marginal likelihood function leads to

$$\begin{aligned} -K(\beta, h : \{\xi, \tau\}) = \log \mathrm{tr}_{\{\sigma\}} \exp &\left[-h \sum_i (\sigma_i - \tau_i)^2 - (\beta/2N) \sum_{ij} (\sigma_i - \sigma_j)^2\right] \\ &+ N\beta m_1^2 - N \log \mathrm{tr}_\sigma \, \mathrm{e}^{2\beta m_1 \sigma - \beta\sigma^2} + \frac{N}{2} \log(h/\pi). \end{aligned} \tag{13}$$

We should keep in mind that the marginal likelihood function depends on $\{\xi\}$ through $\{\tau\}$ by equation (3). Obviously, the above marginal likelihood function $-K(\beta, h : \{\xi, \tau\})$ depends on the data set $\{\xi, \tau\}$ (*quenched disorder* in the context of spin systems). We should average this marginal likelihood function over the distribution of $\{\xi, \tau\}$ to investigate the *averaged case performance* of the maximum marginal likelihood method. In the next section, we carry out this average explicitly.

4. Averaged case performance

In order to investigate the averaged case performance of the maximum marginal likelihood method, we should average the marginal likelihood function $-K(\beta, h : \{\xi, \tau\})$ with respect to $\{\xi\}$ and $\{\tau\}$. Then, we define the averaged marginal likelihood function as

$$\begin{aligned} -K(\beta, h) &\equiv [-K(\beta, h : \{\xi, \tau\})]_{\{\xi,\tau\}} \\ &= -\mathrm{tr}_{\{\xi\}} \int_{-\infty}^{\infty} \{\mathrm{d}\tau\} P_{\beta_s}(\{\xi\}) P_{a_0,a_\tau}(\{\tau\}|\{\xi\}) K(\beta, h : \{\xi, \tau\}) \end{aligned} \tag{14}$$

where we defined $\{\mathrm{d}\tau\} \equiv \mathrm{d}\tau_1\, \mathrm{d}\tau_2 \cdots \mathrm{d}\tau_N$.

To carry out the average, we first rewrite the product of the distribution of the source image and the noise channel as follows:

$$\begin{aligned} P_{\beta_s,a_0,a_\tau}(\{\xi\}, \{\tau\}) &= P_{\beta_s}(\{\xi\}) P_{a_0,a_\tau}(\{\tau\}|\{\xi\}) \\ &= \prod_i \left(\frac{\mathrm{e}^{2\beta_s m_0 \xi_i - \beta_s \xi_i^2}}{\mathrm{tr}_{\xi_i}\, \mathrm{e}^{2\beta_s m_0 \xi_i - \beta_s \xi_i^2}} \right) \frac{\exp\left[-\frac{1}{2a_\tau^2}(\tau_i - a_0 \xi_i)^2\right]}{\sqrt{2\pi} a_\tau} \equiv \prod_i P_{\beta_s,a_0,a_\tau}(\xi_i, \tau_i) \end{aligned} \tag{15}$$

where m_0 should obey the following saddle point equation in the limit of $N \to \infty$:

$$m_0 = \frac{\mathrm{tr}_\xi \xi \exp(2m_0\beta_s\xi - \beta_s\xi^2)}{\mathrm{tr}_\xi \exp(2m_0\beta_s\xi - \beta_s\xi^2)}. \tag{16}$$

Then, in the limit of $N \to \infty$, the first term of the marginal likelihood function (13) becomes the *self-average quantity* and leads to

$$\begin{aligned} &\log \mathrm{tr}_{\{\sigma\}} \exp\left[-h \sum_i (\sigma_i - \tau_i)^2 - (\beta/2N) \sum_{ij} (\sigma_i - \sigma_j)^2 \right] \\ &\quad = -N\beta m^2 + N \cdot \frac{1}{N} \sum_i \log \mathrm{tr}_{\sigma_i}\, \mathrm{e}^{-h(\sigma_i - \tau_i)^2 + 2\beta m \sigma_i - \beta \sigma_i^2} \\ &\quad = -N\beta m^2 + N\, \mathrm{tr}_\xi \int_{-\infty}^{\infty} \mathrm{d}\tau\, P_{\beta_s,a_0,a_\tau}(\xi, \tau) \log \mathrm{tr}_\sigma\, \mathrm{e}^{-h(\sigma-\tau)^2 + 2\beta m\sigma - \beta\sigma^2} \\ &\quad = -N\beta m^2 + N \frac{\mathrm{tr}_\xi\, \mathrm{e}^{2\beta_s m_0 \xi - \beta_s \xi^2}}{Z_0(\beta_s)} \int_{-\infty}^{\infty} Dx \log \mathrm{tr}_\sigma\, \mathrm{e}^{-h(\sigma - a_\tau x - a_0 \xi)^2 + 2\beta m\sigma - \beta\sigma^2} \\ &\quad \equiv [\log Z(\beta, h)]_{\{\xi,\tau\}} \end{aligned} \tag{17}$$

where Dx means the Gaussian integral measure defined by $Dx \equiv \mathrm{d}x\, \mathrm{e}^{-x^2/2} \big/ \sqrt{2\pi}$ and magnetization m obeys the following saddle point equation:

$$m = \frac{\mathrm{tr}_\xi\, \mathrm{e}^{2\beta_s m_0 \xi - \beta_s \xi^2}}{Z_0(\beta_s)} \int_{-\infty}^{\infty} Dx \left[\frac{\mathrm{tr}_\sigma \sigma\, \mathrm{e}^{-(h+\beta)\sigma^2 + 2(ha_\tau x + a_0 h\xi + \beta m)\sigma}}{\mathrm{tr}_\sigma\, \mathrm{e}^{-(h+\beta)\sigma^2 + 2(ha_\tau x + a_0 h\xi + \beta m)\sigma}} \right]. \tag{18}$$

Finally, we obtain the average marginal likelihood per pixel (14) as follows:

$$\begin{aligned} L(\beta, h) \equiv -\frac{K(\beta, h)}{N} &= -\beta m^2 + \frac{\mathrm{tr}_\xi\, \mathrm{e}^{2\beta_s m_0 \xi - \beta_s \xi^2}}{Z_0(\beta_s)} \int_{-\infty}^{\infty} Dx \log \mathrm{tr}_\sigma\, \mathrm{e}^{-h(\sigma - a_\tau x - a_0 \xi)^2 + 2\beta m\sigma - \beta\sigma^2} \\ &\quad + \beta m_1^2 - \log \mathrm{tr}_\sigma\, \mathrm{e}^{2\beta m_1 \sigma - \beta\sigma^2} + \frac{1}{2} \log(h/\pi). \end{aligned} \tag{19}$$

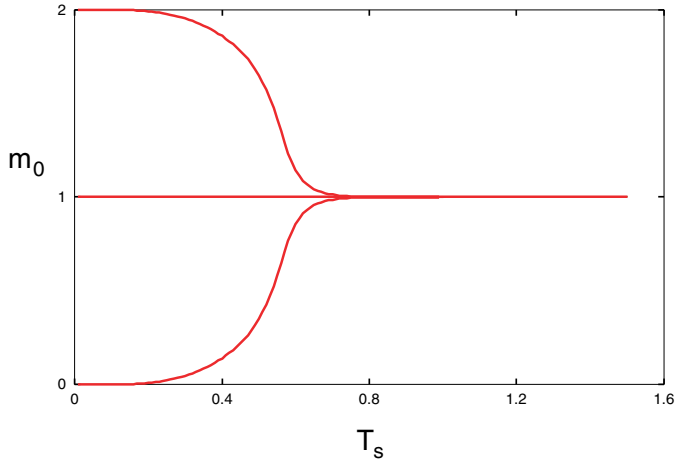

Figure 1. The source magnetization m_0 as a function of $T_s = \beta_s^{-1}$. At the ground state, three states 0, 1, 2 degenerate. For $T_s > 0$, the middle state $m_0 = 1$ becomes a globally stable state.

In the next two subsections, we investigate the hyper-parameter dependence of the averaged marginal likelihood function we obtained here explicitly for the cases $\mathcal{Q} = 3$ and 4.

4.1. Analysis of the $\mathcal{Q} = 3$ case

In this subsection, we consider the case $\mathcal{Q} = 3$, namely, each pixel takes $\xi, \sigma = 0, 1, 2$. In order to investigate the hyper-parameter dependence of the above averaged marginal likelihood function, we first consider the magnetization of the original image. For the $\mathcal{Q} = 3$ case, the saddle point equation with respect to magnetization m_0 of the original image leads to

$$m_0 = \frac{\mathrm{e}^{2m_0\beta_s - \beta_s} + 2\,\mathrm{e}^{4m_0\beta_s - 4\beta_s}}{1 + \mathrm{e}^{2m_0\beta_s - \beta_s} + \mathrm{e}^{4m_0\beta_s - 4\beta_s}}. \tag{20}$$

We plot the magnetization m_0 as a function of temperature $T_s = \beta_s^{-1}$ in figure 1. From this figure, we find that at $T_s = 0$, the three states, namely, $m_0 = 0, 1, 2$, degenerate and the values of the corresponding free energies are all the same. However, when $T_s > 0$, the middle state $m_0 = 1$ has the lowest free energy. As T_s increases, the system goes to the paramagnetic phase. The magnetization m_0 in the paramagnetic state is $m_0(\mathrm{para}) = (0 + 1 + 2)/3 = 1$. For given original images whose magnetization is m_0 at temperature T_s, we evaluate the mean square error (8).

4.1.1. Results: mean square error. In figure 2, we plot the mean square error D as a function of $T = \beta^{-1}$. We set $a_\tau = a_0 = 1$ and set h to its optimal value $h_{\mathrm{opt}} = 1/2$ so as to satisfy the condition

$$-\frac{1}{2a_\tau^2}\sum_i (\tau_i - a_0\xi_i)^2 = -h_{\mathrm{opt}}\sum_i (\tau_i - \sigma_i)^2 \tag{21}$$

for arbitrary configurations of $\{\xi\}$ and $\{\sigma\}$. From these figures, we find that the mean square error takes its minimum at the true hyper-parameter value, namely, $T = T_s = 0.75$ in the context of the MPM estimation [7].

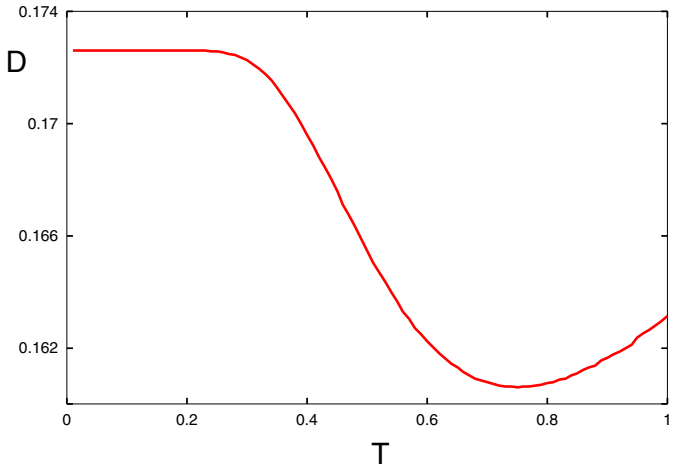

Figure 2. Mean square error D as a function of temperature T. We set the source temperature to $T_s = 0.75$ and choose noise level $a_\tau = a_0 = 1$. We also set field h to its optimal value $h = 1/2$. We find that the mean square error takes its minimum at $T = T_s$.

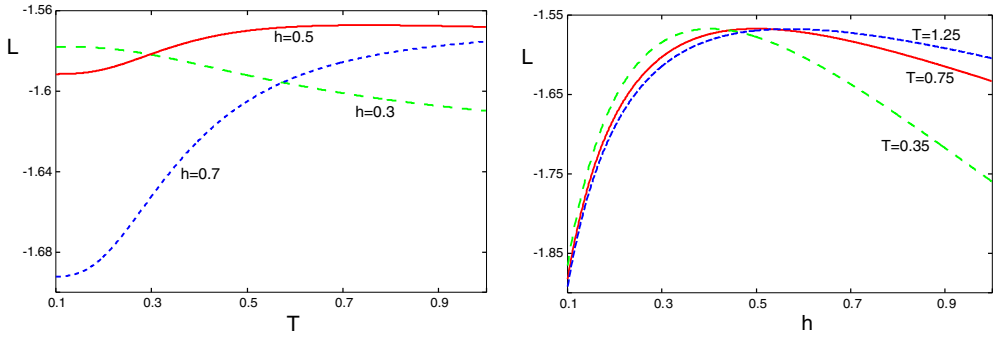

Figure 3. Averaged marginal likelihood L as a function of T (left) and h (right). We set $T_s = 0.75$, $a_\tau = a_0 = 1$. The values of magnetization m_1, m are used as solutions of equations (12) and (18). These figures show that the averaged marginal likelihood function takes its maximum at $(T, h) = (T_s, h_{\rm opt}) = (0.75, 0.5)$.

4.1.2. Results: averaged marginal likelihood function. In practice, we need to estimate these hyper-parameter values from a given degraded image $\{\tau\}$. Therefore, for this model system, we next consider the hyper-parameter estimation in the context of maximization of the marginal likelihood function (for the Ising case, see [9]). In figure 3, we plot the averaged marginal likelihood L as a function of T and h. From these two figures, we find that the averaged marginal likelihood function is maximized at the true hyper-parameter values, namely, $T = T_s = 0.75$ and $h = h_{\rm opt} = 1/2$.

4.2. Analysis of the $\mathcal{Q} = 4$ case

We next show the result for the case $\mathcal{Q} = 4$. The source magnetization m_0 as a function of temperature T_s

$$m_0 = \frac{\mathrm{e}^{2m_0\beta_s-\beta_s} + 2\,\mathrm{e}^{4m_0\beta_s-4\beta_s} + 3\,\mathrm{e}^{6m_0\beta_s-9\beta_s}}{1 + \mathrm{e}^{2m_0\beta_s-\beta_s} + \mathrm{e}^{4m_0\beta_s-4\beta_s} + \mathrm{e}^{6m_0\beta_s-9\beta_s}} \tag{22}$$

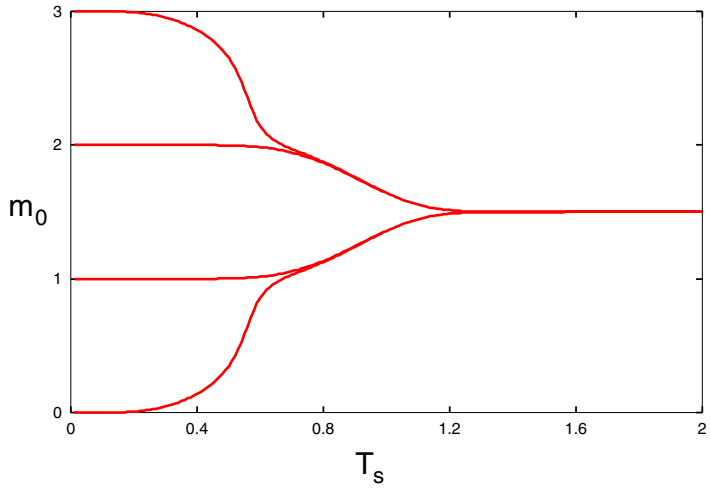

Figure 4. The source magnetization for the case $\mathcal{Q} = 4$. At $T_s = 0$, four states $m_0 = 0, 1, 2$ and 3 degenerate. For finite temperature $T_s > 0$, the middle two states $m_0 = 1, 2$ become globally stable states. As temperature increases, the system goes to paramagnetic phase specified by the magnetization $m_0(\text{para}) = (0 + 1 + 2 + 3)/4 = 1.5$.

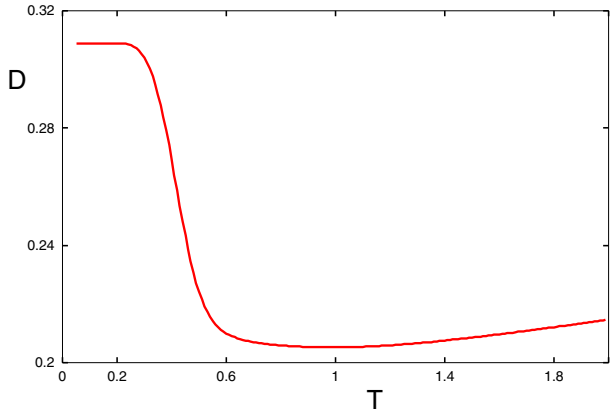

Figure 5. The mean square error D (right) as a function of T for the case $\mathcal{Q} = 4$. The mean square error takes its minimum at $T = T_s = 1.0$. We set $a_\tau = a_0 = 1$ and $h = h_{\text{opt}} = 1/2$.

is shown in figure 4. In this figure, we find that at the ground state, four states $m_0 = 0, 1, 2$ and 3 degenerate, however, $T_s > 0$, the middle two states $m_0 = 1$ and $m_0 = 2$ become globally stable states. As T_s increases, the system goes to paramagnetic phase whose magnetization is $m_0(\text{para}) = (0 + 1 + 2 + 3)/4 = 1.5$.

To use original images which have the magnetization m_0, we set the temperature $T_s = 1.0$ and select the images at this temperature.

4.2.1. Results: mean square error. We evaluate the mean square error D. In figure 5, we plot the temperature T dependence of the magnetization m and corresponding mean square error. From the figure, we find that the mean square error takes its minimum at $T = T_s$.

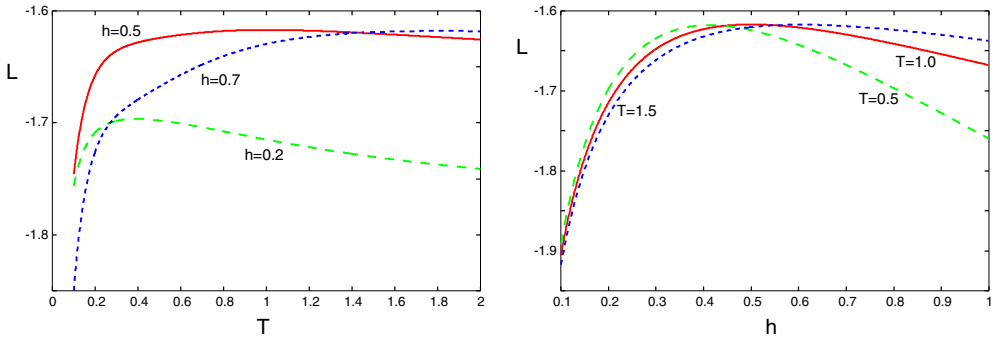

Figure 6. The marginal likelihood L as a function of T (left) and h (right) for the case $\mathcal{Q} = 4$. In the left-hand figure, we set the value of h to its optimal value $1/2$ for $a_\tau = a_0 = 1$. In the right-hand figure, we set $T_d = T_s = 1.0$.

4.2.2. Results: marginal likelihood function. We next consider the hyper-parameter dependence of the averaged marginal likelihood function. In figure 6, we see that the marginal likelihood takes its maximum at the true values of the hyper-parameters.

From the results obtained in this section, we conclude that the criterion of maximizing the marginal likelihood function works well to obtain the optimal hyper-parameters. Therefore, our problem is now how we maximize the marginal likelihood function for a given degraded image. The simplest strategy to attempt it is to construct gradient descent of the marginal likelihood function. For this strategy, we usually obtain a kind of Boltzmann machine learning equations and their properties are investigated by Inoue and Tanaka [9] in the context of the black and white image restoration. Besides the gradient descent, the *EM algorithm (expectation maximum algorithm)* [10] is well known in the field of statistics. In the next section, we investigate the averaged performance of the EM algorithm for determination of the optimal hyper-parameters in Bayesian grey scale image restoration.

5. EM algorithm

In the previous section, we found that the data-averaged marginal likelihood function takes its maximum at the true values of the hyper-parameters. Therefore, the next problem to be tackled by us is how one maximizes the likelihood function with respect to the hyper-parameters. Usually, the direct maximization via, for example, the gradient descent method or some other optimization techniques requires enormous computational costs.

To make the problem much more tractable, the EM algorithm (expectation maximum algorithm) [10] is widely used. For black and white image restoration, Inoue and Tanaka [9] investigated the performance of the EM algorithm and gradient descent to maximize the marginal likelihood function by using the infinite range random field Ising model. They compared these two methods and found that the EM algorithm shows faster convergence to the solution than the gradient descent does. Here, we investigate properties of the EM algorithm for grey scale image restoration by using the $\mathcal{Q}$-Ising model.

The EM algorithm maximizes the following Q *function*, namely, the expectation of the log-likelihood function $\log P_h(\{\tau\}|\{\sigma\})P_\beta(\{\sigma\})$ over the time-dependent posterior distribution

$P_{\beta^{(t)},h^{(t)}}(\{\sigma\}|\{\tau\})$ at each iteration step t

$$
\begin{aligned}
Q(\beta,h|\beta^{(t)},h^{(t)}:\{\xi,\tau\}) &= \mathrm{tr}_{\{\sigma\}} P_{\beta^{(t)},h^{(t)}}(\{\sigma\}|\{\tau\}) \log P_h(\{\tau\}|\{\sigma\}) P_\beta(\{\sigma\}) \\
&= -h\left[\frac{\mathrm{tr}_{\{\sigma\}} \sum_i (\sigma_i-\tau_i)^2 \,\mathrm{e}^{-h^{(t)}\sum_i(\sigma_i-\tau_i)^2-(\beta^{(t)}/2N)\sum_{ij}(\sigma_i-\sigma_j)^2}}{\mathrm{tr}_{\{\sigma\}} \,\mathrm{e}^{-h^{(t)}\sum_i(\sigma_i-\tau_i)^2-(\beta^{(t)}/2N)\sum_{ij}(\sigma_i-\sigma_j)^2}}\right] \\
&\quad -\frac{\beta}{2N}\left[\frac{\mathrm{tr}_{\{\sigma\}} \sum_{ij} (\sigma_i-\sigma_j)^2 \,\mathrm{e}^{-h^{(t)}\sum_i(\sigma_i-\tau_i)^2-(\beta^{(t)}/2N)\sum_{ij}(\sigma_i-\sigma_j)^2}}{\mathrm{tr}_{\{\sigma\}} \,\mathrm{e}^{-h^{(t)}\sum_i(\sigma_i-\tau_i)^2-(\beta^{(t)}/2N)\sum_{ij}(\sigma_i-\sigma_j)^2}}\right] \\
&\quad -\log Z_\Pi(\beta) - \log Z_L(h) \\
&= h\frac{\partial}{\partial h^{(t)}} \log Z(\beta^{(t)},h^{(t)}) + \frac{\beta}{2N}\frac{\partial}{\partial(\beta^{(t)}/2N)} \log Z(\beta^{(t)},h^{(t)}) \\
&\quad + N\beta m_1^2 - N\log \mathrm{tr}_\sigma \,\mathrm{e}^{2\beta m_1\sigma-\beta\sigma^2} + \frac{N}{2}\log(\pi/h) \qquad (23)
\end{aligned}
$$

where $Z(\beta^{(t)},h^{(t)})$ is a partition function given by

$$
Z(\beta^{(t)},h^{(t)}) = \mathrm{tr}_{\{\sigma\}} \,\mathrm{e}^{-h^{(t)}\sum_i(\sigma_i-\tau_i)^2-(\beta^{(t)}/2N)\sum_{ij}(\sigma_i-\sigma_j)^2}. \qquad (24)
$$

The above Q function is regarded as *energy* in contrast to the marginal likelihood function as *free energy*. The EM algorithm maximizes the marginal likelihood function indirectly and it guarantees local maximum solutions of the marginal likelihood function. It is important for us to bear in mind that the parameters to be maximized, namely, β and h, do not appear in the posterior distribution. Thus, the expectations of the quantities $\sum_i(\tau_i-\sigma_i)^2$ or $\sum_{ij}(\sigma_i-\sigma_j)^2$ over the posterior distribution become linear with respect to β and h, and as a result, the maximization conditions lead to simple non-linear maps. This is one of the advantages of the EM algorithm and this fact makes the optimization problem more tractable.

As our interest here is the averaged case performance of the EM algorithm instead of the performance for specific data sets $\{\xi,\tau\}$, we should evaluate the averaged Q function, that is to say,

$$
\begin{aligned}
Q(\beta,h|\beta^{(t)},h^{(t)}) &= [Q(\beta,h|\beta^{(t)},h^{(t)}:\{\xi,\tau\})]_{\{\xi,\tau\}} \\
&= h\frac{\partial}{\partial h^{(t)}}[\log Z(\beta^{(t)},h^{(t)})]_{\{\xi,\tau\}} + \beta\frac{\partial}{\partial \beta^{(t)}}[\log Z(\beta^{(t)},h^{(t)})]_{\{\xi,\tau\}} \\
&\quad + N\beta m_1^2 - N\log \mathrm{tr}_\sigma \,\mathrm{e}^{2\beta m_1\sigma-\beta\sigma^2} + \frac{N}{2}\log(\pi/h). \qquad (25)
\end{aligned}
$$

After substituting the average $[\log Z(\beta^{(t)},h^{(t)})]_{\{\xi,\tau\}}$ (see (17)) into the above expression, we obtain the data-averaged Q function per pixel as follows:

$$
\begin{aligned}
q(\beta,h|\beta^{(t)},h^{(t)}) &\equiv \frac{Q(\beta,h|\beta^{(t)},h^{(t)})}{N} = -h\frac{\mathrm{tr}_\xi \,\mathrm{e}^{2\beta_s m_0\xi-\beta_d\xi^2}}{Z_0(\beta_s)} \\
&\quad \times \int_{-\infty}^{\infty} Dx \left[\frac{\mathrm{tr}_\sigma(\sigma-a_\tau x-a_0\xi)^2 \,\mathrm{e}^{-h^{(t)}(\sigma-a_\tau x-a_0\xi)^2+2\beta^{(t)}m\sigma-\beta^{(t)}\sigma^2}}{\mathrm{tr}_\sigma \,\mathrm{e}^{-h^{(t)}(\sigma-a_\tau x-a_0\xi)^2+2\beta^{(t)}m\sigma-\beta^{(t)}\sigma^2}}\right] \\
&\quad -\beta m^2 + \beta\frac{\mathrm{tr}_\xi \,\mathrm{e}^{2\beta_s m_0\xi-\beta_s\xi^2}}{Z_0(\beta_s)} \\
&\quad \times \int_{-\infty}^{\infty} Dx \left[\frac{\mathrm{tr}_\sigma(2m\sigma-\sigma^2) \,\mathrm{e}^{-h^{(t)}(\sigma-a_\tau x-a_0\xi)^2+2\beta^{(t)}m\sigma-\beta^{(t)}\sigma^2}}{\mathrm{tr}_\sigma \,\mathrm{e}^{-h^{(t)}(\sigma-a_\tau x-a_0\xi)^2+2\beta^{(t)}m\sigma-\beta^{(t)}\sigma^2}}\right] \\
&\quad + \beta m_1^2 - \log \mathrm{tr}_\sigma \,\mathrm{e}^{2\beta m_1\sigma-\beta\sigma^2} + \frac{1}{2}\log(\pi/h). \qquad (26)
\end{aligned}
$$

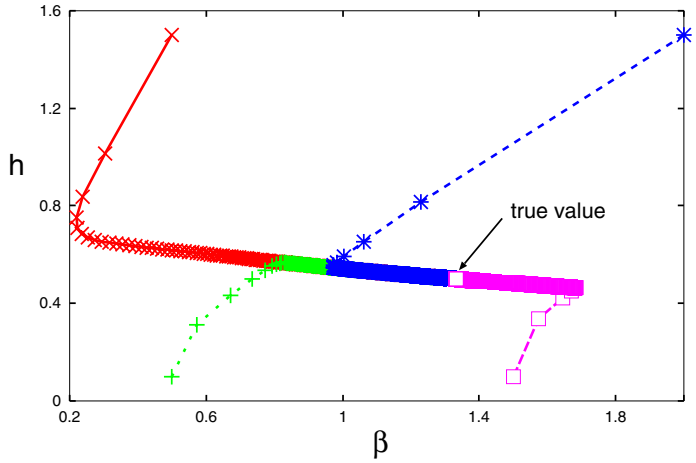

Figure 7. Flows of hyper-parameters (β, h) by using the EM algorithm. The noise level and source temperature are selected as $a_\tau = a_0 = 1$ and $T_s = 0.75$, respectively. Therefore, the true values of β and h are $\beta = 1/T_s = 1.33$ and $h = h_{\rm opt} = 1/2$. We find that all flows converge to the correct solution (1.33, 0.5).

For this data-averaged Q function, the EM algorithm updates the hyper-parameters β and h according to the following rules:

$$\beta^{(t+1)} = \arg\max_{\beta} q(\beta, h|\beta^{(t)}, h^{(t)}) \tag{27a}$$

$$h^{(t+1)} = \arg\max_{h} q(\beta, h|\beta^{(t)}, h^{(t)}). \tag{27b}$$

After calculating the above non-linear maps, we obtain the maximum likelihood estimate as a fixed point of the dynamics. In the next two subsections, we investigate the performance of these iterations for the cases $\mathcal{Q} = 3$ and 4.

5.1. Analysis of the $\mathcal{Q} = 3$ case

To demonstrate the EM algorithm to infer the hyper-parameters β and h, we first consider the case $\mathcal{Q} = 3$. It is important for us to keep in mind that the magnetizations m and m_1 appearing in the equations (27*a*) and (27*b*) are used as equilibrium values, namely, for each time step, m and m_1 should satisfy the saddle point equations (12) and (18) with $(\beta, h) = (\beta^{(t)}, h^{(t)})$, respectively. In figure 7, we plot several hyper-parameter flows calculated by the EM algorithm (27*a*) and (27*b*). We found that the EM algorithm achieves convergence to the correct solution even if we start the algorithm from any point which is far from the solution. However, the speed of the convergence depends on the initial condition of the non-linear EM iterations (27*a*) and (27*b*). We plot the time dependence of the hyper-parameters β and h in figure 8. This figure shows that when we choose large β as a starting point of maps (27*a*) and (27*b*), in other words, if we start the EM updates in the ferromagnetic state at low temperature, the speed of convergence becomes very slow. In figure 9, we also plot the time-dependent Q function $Q(t : \beta, h) \equiv q(\beta, h|\beta^{(t)}, h^{(t)})$ as a function of β and h provided that the tth values of the hyper-parameters $\beta^{(t)}$ and $h^{(t)}$ are given. We plot for the cases $t = 0, 1, 100$ and $t = 1000$

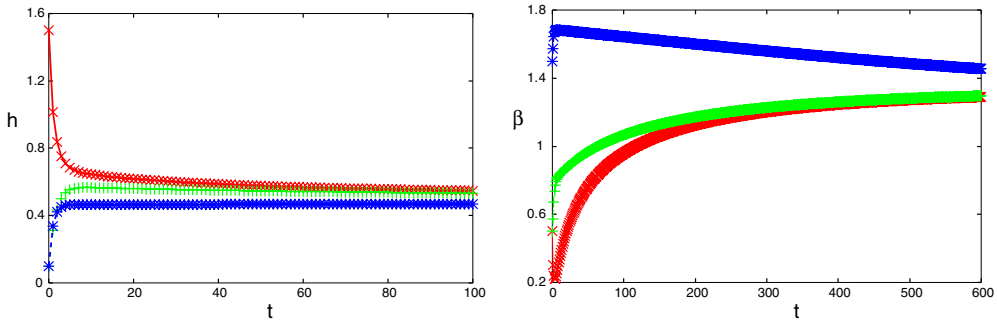

Figure 8. Time evolutions of hyper-parameters h (left) and β (right). When we choose large β as an initial state, the speed of convergence becomes very slow.

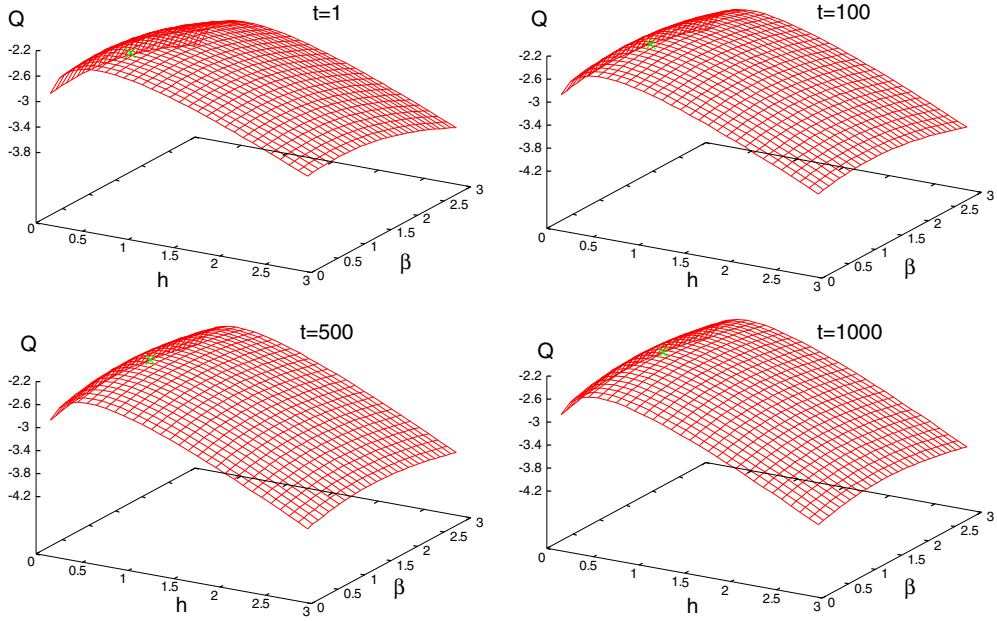

Figure 9. The time evolution of the surface of the Q function. From the upper left to the lower right, $t = 0$, 100, 500 and $t = 1000$ cases are plotted. We set $\beta = 2.0$, $h = 1.0$ as the initial state of the evolution of the hyper-parameters.

time steps. From this figure, we see that at an early stage of the EM update, the slope with respect to h for a given β is much more steep than that of β. This property means that the EM algorithm shows fast convergence to the solution of $h = h_{\rm opt} = 1/2$ at the beginning of the EM update. After that, a little movement in the β-direction starts to converge to the solution. The speed of the convergence is very slow. This is because the slope with respect to β in the surface of the Q function is relatively slack. The time evolutions of data-averaged marginal likelihood function during the EM algorithm are shown in figure 10. This figure shows that

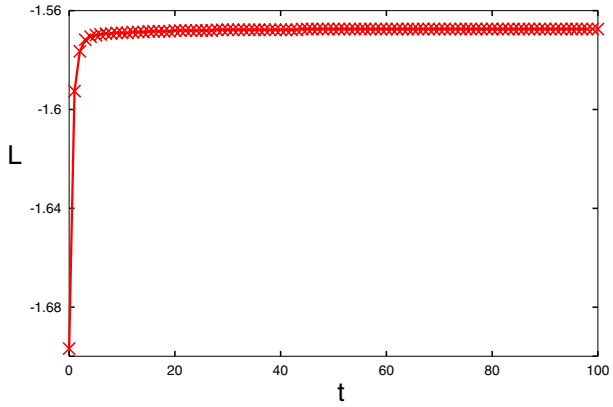

Figure 10. The time evolutions of the data-averaged marginal likelihood function during the EM algorithm. We set the initial values of the hyper-parameters $(\beta^{(0)}, h^{(0)}) = (0.5, 1.5)$.

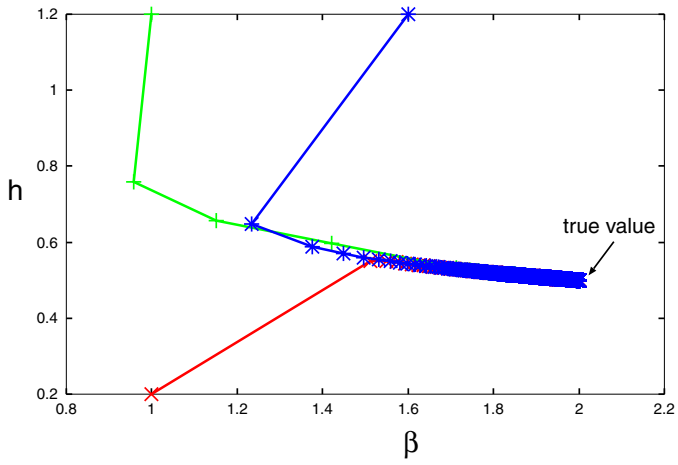

Figure 11. Flows of hyper-parameters $\beta^{(t)}$ and $h^{(t)}$ for the case $\mathcal{Q} = 4$. We set the true values of the hyper-parameters $\beta = 1/T_s = 2.0$ and $h = h_{\rm opt} = 1/2$. We find that all flows converge to the true point.

the data-averaged marginal likelihood function increases monotonically during EM updates (27*a*) and (27*b*) and converges to its global maximum.

5.2. Analysis of the $\mathcal{Q} = 4$ case

In order to investigate the effects of the number of the grey scale levels $\mathcal{Q}$ on the dynamics of the hyper-parameters, we briefly show the results for the case $\mathcal{Q} = 4$. We simply show the flows and the time evolutions of the hyper-parameters in figures 11 and 12. We find that the flows converge to the true value and the final solution is independent of the choice of the initial condition, however, the speed of the convergence becomes very slow. For this $\mathcal{Q} = 4$ case, the convergence of the h-direction is much faster than that of the β-direction as we saw in the case $\mathcal{Q} = 3$.

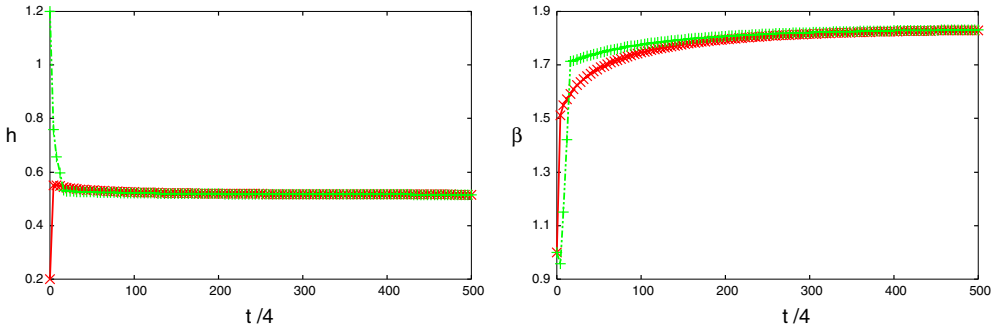

Figure 12. Time developments of the hyper-parameters h (left) and β (right) for the case $\mathcal{Q} = 4$. We should keep in mind that the time axis is scaled by $1/4$, namely, the actual range of time t is $[0 : 2000]$.

6. Summary

In this paper, we investigated the dynamics of the hyper-parameter estimation via the EM algorithm by analysis of the infinite range $\mathcal{Q}$-Ising model. We calculated the data-average marginal likelihood function and found that it takes a maximum at the true values of the hyper-parameters. The EM algorithm was demonstrated for our model system to obtain the maximum likelihood estimate of the hyper-parameters. We evaluated the time evolutions of the hyper-parameters and the surface of the Q function. We found from the time evolutions of the surface of the Q function that at the early stage of the EM update, the slope of the Q function with respect to h for a given β is much steeper than that of β. This property means that the EM algorithm shows fast convergence along the h-direction at the initial stage of the EM updates, however, the movement of the β-direction is extremely slow. This is because the slope with respect to β in the surface of the Q function is relatively slack. From the viewpoint of computational cost, it was revealed that the increase of the grey scale levels $\mathcal{Q}$ is serious to obtain the solution. We hope our analysis gives an insight to improve the slow convergence which is the nature of the EM algorithm for the hyper-parameters estimation in the Bayesian grey scale image restoration.

References

[1] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: an Introduction* (Oxford: Oxford University Press)
[2] Tanaka K 2002 *J. Phys.* A **35** R81
[3] Pryce J M and Bruce A D 1995 *J. Phys.* A **28** 511
[4] Nishimori H and Wong K Y M 1999 *Phys. Rev.* E **60** 132
[5] Marroquin J, Mitter S and Poggio T 1987 *J. Am. Stat. Assoc.* **82** 76
[6] Carlucci D M and Inoue J 1999 *Phys. Rev.* E **60** 2543
[7] Inoue J and Carlucci D M 2001 *Phys. Rev.* E **64** 036121
[8] Bollé D, Rieger H and Shim G M 1994 *J. Phys. A: Math. Gen.* **27** 3411
[9] Inoue J and Tanaka K 2002 *Phys. Rev.* E **65** 016125
[10] Dempster A P, Larid N M and Rubin D B 1977 *J. Roy. Soc. Stat.* B **39** 1