Journal of Physics: Conference Series **473** (2013) 012006 doi:10.1088/1742-6596/473/1/012006

A generalization of improved susceptibility propagation

M Yasuda

Graduate School of Science and Engineering, Yamagata University, Yonezawa 992-8510, Yamagata pref., Japan

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

Abstract. Recently, an effective susceptibility propagation method for binary Markov random fields based on a concept of diagonal consistency was proposed. This improved susceptibility propagation is a powerful method and exhibits a robust performance for various types of network structures. In this paper, a generalization of the improved susceptibility propagation using an orthonormal function expansion is proposed in which any pairwise potential functions and multi-valued random variables are acceptable. In the latter part of this paper, the proposed method is applied to a direct problem and an inverse problem on a generalized sparse prior, which is a recently proposed prior model for natural images.

1. Introduction

There is an increasing demand in the new interdisciplinary scientific fields that involve both computer sciences and statistical physics [1, 2] for techniques that can be used to evaluate local statistical quantities such as local magnetizations and local susceptibilities (covariances) of Markov random fields (MRFs) with finite sizes.

Belief propagations (BPs) are one of the most popular message-passing types of algorithms that are used to compute approximately the local magnetizations of MRFs [3]; they are equivalent to the Bethe approximations [4] used in statistical physics [5, 6]. In the last decade, a suitable message-passing technique called susceptibility propagation (SusP) (also called variational linear response in the field of information sciences) was developed to compute the local susceptibilities of MRFs [7–10]. In general, SusPs are constructed by combining BPs and linear response methods.

Recently, the author presented an effective SusP for Ising models, referred to as *improved susceptibility propagation* (I-SusP) using a concept of *diagonal consistency* [11]. Briefly, the concept of diagonal consistency is as follows. In Ising models, the second-order moment, $\langle x_i^2 \rangle$, are trivially one because $x_i \in \{+1, -1\}$. However, the variances obtained by SusPs are generally not equal to one due to higher-order effects being neglected [12]. I-SusPs are obtained by overcoming such a diagonal inconsistency problem. I-SusPs are reduced to normal SusPs on tree systems and include the adaptive Thouless-Anderson-Palmer equation [13, 14] as a special case. Almost simultaneously, a similar investigation of Ising models was presented by Raymond and Ricci-Tersenghi [15], who considered not only the diagonal consistency but also an off-diagonal consistency [16].

The I-SusP is a powerful method and exhibits a robust performance for various types of network structures. However, the original I-SusP can be applied to Ising models only. Thus, in the present paper, a generalization scheme of I-SusP using an orthonormal function expansion that was successfully used in a Bethe approximation for a general pairwise MRF [17] is proposed, in which any pairwise potential functions and multi-valued random variables are acceptable.

The remainder of this paper is organized as follows. In section 2, a pairwise MRF and its orthogonal function expansion using Chebyshev polynomials are introduced. In section 3, a Gibbs free energy for the expanded pairwise MRF derived in the previous section and its diagonal-block consistency are shown. The diagonal-block consistency plays the central role in the proposed method. In section 4, by using the diagonal-block consistency, an I-SusP based on the Bethe approximation for the expanded pairwise MRF is proposed. In section 5, the proposed method is applied to *generalized sparse priors* and is numerically verified by some results of numerical experiments. Generalized sparse priors are prior models for natural images proposed in [18]. Finally, section 6 concludes the paper with some remarks.

2. Discrete Pairwise MRF and Orthogonal Function Expansion

Let us consider an undirected graph $G = G(V, E)$ with n nodes, where $V = \{1, 2, \ldots, n\}$ is the set of all nodes and $E \subset V \times V$ is the set of all undirected links (i, j) except self-connecting links (i, i). Note that, although links (i, j) and (j, i) indicate the same undirected link, the set E has both (i, j) and (j, i) as its elements. On the graph G, let us consider a pairwise MRF with discrete random variables $\boldsymbol{x} = \{x_i \in \{0, 1, \ldots, q-1\} \mid i \in V\}$ expressed by

$$P_G(\boldsymbol{x}) \propto \exp\Big(\sum_{i\in V} \theta_i(x_i) + \frac{1}{2}\sum_{(i,j)\in E} w_{(i,j)}(x_i, x_j)\Big), \tag{1}$$

where the expressions $\{\theta_i(x_i)\}$ and $\{w_{(i,j)}(x_i, x_j)\}$ are potential functions of nodes and links, respectively. Because the graph is an undirected graph, relations $w_{(i,j)}(x_i, x_j) = w_{(j,i)}(x_j, x_i)$ are assumed. Relations $w_{(i,j)}(x_i, x_j) = w_{(i,j)}(x_j, x_i)$ are not enforced, that is, $w_{(i,j)}(x_i, x_j)$ does not have to be a symmetric function with respect to x_i and x_j.

Let us introduce an orthonormal set $\{\phi_k(x) \in \mathbb{R} \mid k = 0, 1, 2, \ldots, q-1\}$ such that

$$\sum_{x=0}^{q-1} \phi_k(x)\phi_l(x) = \delta_{k,l}, \tag{2}$$

where $\delta_{k,l}$ is the Kronecker delta. If the orthonormal set is a basis, we can expand a function $f(x)$ of an integer variable $x \in \{0, 1, \ldots, q-1\}$ by using the orthonormal set as

$$f(x) = \sum_{k=0}^{q-1} \alpha_k \phi_k(x), \quad \text{where} \quad \alpha_k := \sum_{x=0}^{q-1} f(x)\phi_k(x),$$

One possible choice of orthonormal set is the set of Chebyshev polynomials given in Appendix A. We use the Chebyshev polynomials as the orthonormal set in this paper.

Using the orthonormal set of polynomials, we can expand the potential functions $\theta_i(x_i)$ and $w_{(i,j)}(x_i, x_j)$ as

$$\theta_i(x_i) = \sum_{k=0}^{q-1} u_{[i,k]}\phi_k(x_i) \quad \text{and} \quad w_{(i,j)}(x_i, x_j) = \sum_{k,l=0}^{q-1} J_{[i,k],[j,l]}\phi_k(x_i)\phi_l(x_j), \tag{3}$$

respectively, where $u_{[i,k]}$ and $J_{[i,k],[j,l]}$ are constants given by

$$u_{[i,k]} := \sum_{x_i=0}^{q-1} \theta_i(x_i)\phi_k(x_i) \quad \text{and} \quad J_{[i,k],[j,l]} := \sum_{x_i,x_j=0}^{q-1} w_{(i,j)}(x_i,x_j)\phi_k(x_i)\phi_l(x_j), \tag{4}$$

respectively. When the functional forms of the potential functions have been specified, the values of the constants are identified uniquely via (4). By these definitions, $J_{[i,k],[j,l]} = J_{[j,l],[i,k]}$ are satisfied, but $J_{[i,k],[j,l]} \neq J_{[i,l],[j,k]}$ in general. By using the expansions in (3) and the relation $\phi_0(x) = 1/\sqrt{q}$ (see Appendix A), the MRF in (1) can be rewritten as

$$P_G(\boldsymbol{x}) = P_G(\boldsymbol{x} \mid \boldsymbol{h}, \boldsymbol{J}) = \frac{1}{Z(\boldsymbol{h}, \boldsymbol{J})} \exp\Big(- \mathcal{H}(\boldsymbol{x}; \boldsymbol{h}, \boldsymbol{J})\Big), \tag{5}$$

where

$$\mathcal{H}(\boldsymbol{x}; \boldsymbol{h}, \boldsymbol{J}) := -\sum_{k=1}^{q-1} h_{[i,k]}\phi_k(x_i) - \frac{1}{2}\sum_{(i,j)\in E}\sum_{k,l=1}^{q-1} J_{[i,k],[j,l]}\phi_k(x_i)\phi_l(x_j)$$

and the expression $Z(\boldsymbol{h}, \boldsymbol{J})$ is the partition function. The expression $h_{[i,k]}$ is defined by

$$h_{[i,k]} := u_{[i,k]} + \frac{1}{\sqrt{q}} \sum_{j\in\partial(i)} J_{[i,k],[j,0]} = u_{[i,k]} + \frac{1}{q}\sum_{j\in\partial(i)}\sum_{x_i,x_j=0}^{q-1} w_{(i,j)}(x_i,x_j)\phi_k(x_i).$$

The notation $\partial(i)$ is the set of nodes connecting to node i. In the expression in (5), the constant terms are neglected because they have no influence on the probability. This rewritten expansion of MRF is in the exponential family. Therefore, it can be found that we can obtain a distribution in the exponential family that is equivalent to the pairwise MRF in (1) by using the orthonormal expansion. In the following, we refer the distribution (5) as "the expanded pairwise MRF".

3. Gibbs Free Energy and Diagonal-Block Consistency

In this section, we formulate a Gibbs free energy (GFE) and its diagonal-block consistency for the expanded pairwise MRF. GFE representations and their diagonal-block consistencies are important components for I-SusPs provided in section 4.

3.1. Gibbs Free Energy

Let us introduce a GFE, which is a dual form of the Helmholtz free energy (HFE) defined by $F(\boldsymbol{h}, \boldsymbol{J}) := -\ln Z(\boldsymbol{h}, \boldsymbol{J})$, in the following variational way. Let us define the variational function with respect to a trial distribution $Q(\boldsymbol{x})$:

$$\mathcal{F}[Q] := \sum_{\boldsymbol{x}} \mathcal{H}(\boldsymbol{x})Q(\boldsymbol{x}) + \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \ln Q(\boldsymbol{x}), \tag{6}$$

and consider the minimization of the variational function under the restrictions

$$\sum_{\boldsymbol{x}} Q(\boldsymbol{x}) = 1, \quad \text{and} \quad \sum_{\boldsymbol{x}} \phi_k(x_i)Q(\boldsymbol{x}) = m_{[i,k]}.$$

These are the restrictions for normalizing and expectations, respectively. By using the Lagrange multipliers, the constrained minimization of $\mathcal{F}[Q]$ leads to a GFE as follows:

$$\mathcal{G}(\boldsymbol{m}) := \min_{Q} \operatorname*{extr}_{z,\boldsymbol{\gamma}} \Big\{ \mathcal{F}[Q] - z\Big(\sum_{\boldsymbol{x}} Q(\boldsymbol{x}) - 1\Big) - \sum_{i\in V} \gamma_i \Big(\sum_{\boldsymbol{x}} \phi_k(x_i)Q(\boldsymbol{x}) - m_{[i,k]}\Big)\Big\}$$

$$= -\boldsymbol{h}^{\mathrm{t}}\boldsymbol{m} + \max_{\boldsymbol{\gamma}} \Big(\boldsymbol{\gamma}^{\mathrm{t}}\boldsymbol{m} + F(\boldsymbol{\gamma}, \boldsymbol{J}) \Big), \tag{7}$$

where the notation "extr" denotes the extremum with respect to the assigned parameters and the notation t is the transposition. In the GFE, we regard $\boldsymbol{m} = \{m_{[i,k]} \mid i \in V; k = 1, 2, \ldots, q-1\}$ as the independent variables. Expression $\boldsymbol{\gamma} = \{\gamma_{i,k} \mid i \in V; k = 1, 2, \ldots, q-1\}$ denote the Lagrange multipliers for the constraints for the corresponding expectations. The expression $\mathcal{G}(\boldsymbol{m})$ is the GFE of the present system and is regarded as a dual expression of the HFE by the following reasoning. Suppose that $\boldsymbol{m}^{\dagger}$ are the values of $\boldsymbol{m}$ that minimize the GFE, i.e., $\boldsymbol{m}^{\dagger} = \arg\min_{\boldsymbol{m}} \mathcal{G}(\boldsymbol{m})$, the following relations are satisfied:

$$\mathcal{G}(\boldsymbol{m}^{\dagger}) = F(\boldsymbol{h}, \boldsymbol{J}) \quad \text{and} \quad m^{\dagger}_{[i,k]} = \sum_{\boldsymbol{x}} \phi_k(x_i) P_G(\boldsymbol{x} \mid \boldsymbol{h}, \boldsymbol{J}) =: m^{\mathrm{ex}}_{[i,k]}. \tag{8}$$

This fact can be easily verified as follows. From the minimum condition of (7) with respect to $\boldsymbol{m}$, we obtain

$$-h_{[i,k]} + \gamma^{*}_{[i,k]} = 0, \tag{9}$$

where $\boldsymbol{\gamma}^{*}$ are values of $\boldsymbol{\gamma}$ that satisfy the maximum conditions in (7), i.e., $\boldsymbol{\gamma}^{*} = \arg\max_{\boldsymbol{\gamma}} \big(\boldsymbol{\gamma}^{\mathrm{t}}\boldsymbol{m} + F(\boldsymbol{\gamma}, \boldsymbol{J}) \big)$, and therefore, $\boldsymbol{\gamma}^{*}$ satisfy relations

$$m_{[i,k]} = \sum_{\boldsymbol{x}} \phi_k(x_i) P_G(\boldsymbol{x} \mid \boldsymbol{\gamma}^{*}, \boldsymbol{J}). \tag{10}$$

Using (9) and (10), we arrive at the relations in (8).

3.2. Linear Response Relation and Diagonal-Block Consistency

It is known that the GFE satisfies a linear response relation that is obtained in the following way. By differentiating (10) with respect to $m_{[j,l]}$, we obtain

$$\delta_{i,j}\delta_{k,l} = \sum_{i' \in V} \sum_{r=1}^{q-1} \Big(\langle \phi_k(x_i)\phi_r(x_{i'}) \rangle_{\boldsymbol{\gamma}^{*}} - m_{[i,k]} m_{[i',r]} \Big) \frac{\partial \gamma^{*}_{[i',r]}}{\partial m_{[j,l]}}, \tag{11}$$

where notation $\langle \cdots \rangle_{\gamma^{*}}$ denotes the expectation with respect to distribution $P_G(\boldsymbol{x} \mid \boldsymbol{\gamma}^{*}, \boldsymbol{J})$. Since $\partial^2 \mathcal{G}(\boldsymbol{m}) / \partial m_{[i,k]} \partial m_{[j,l]} = \partial \gamma^{*}_{[i,k]} / \partial m_{[j,l]}$, (11) is collectively expressed by

$$\boldsymbol{H}(\boldsymbol{m})^{-1} = \boldsymbol{\chi}(\boldsymbol{m}), \tag{12}$$

where $n(q-1) \times n(q-1)$ matrices $\boldsymbol{H}(\boldsymbol{m})$ and $\boldsymbol{\chi}(\boldsymbol{m})$ are the Hessian matrix of the GFE and the susceptibility matrix, respectively. They are defined by

$$\big[\boldsymbol{H}(\boldsymbol{m})\big]_{e(i,k),e(j,l)} := \frac{\partial^2 \mathcal{G}(\boldsymbol{m})}{\partial m_{[i,k]} \partial m_{[j,l]}}, \quad \big[\boldsymbol{\chi}(\boldsymbol{m})\big]_{e(i,k),e(j,l)} := \langle \phi_k(x_i)\phi_l(x_j) \rangle_{\boldsymbol{\gamma}^{*}} - m_{[i,k]} m_{[j,l]},$$

where $e(i,k) := (i-1)q + k$ and the notation $[\cdots]_{a,b}$ indicates the (a,b)-element of a matrix. The relation in (12) is known as one type of linear response relation and always holds for any $\boldsymbol{m}$.

By using the relations in (10) and the orthonormal set, one-variable marginal distributions of $P_G(\boldsymbol{x} \mid \boldsymbol{\gamma}^{*}, \boldsymbol{J})$, $P_i(x_i \mid \boldsymbol{\gamma}^{*}, \boldsymbol{J}) := \sum_{\boldsymbol{x} \setminus \{x_i\}} P_G(\boldsymbol{x} \mid \boldsymbol{\gamma}^{*}, \boldsymbol{J})$, can be expanded as $P_i(x_i \mid \boldsymbol{\gamma}^{*}, \boldsymbol{J}) =$

$q^{-1}+\sum_{k=1}^{q-1} m_{[i,k]}\phi_k(x_i)$. Therefore, because $\langle\phi_k(x_i)\phi_l(x_i)\rangle_{\boldsymbol{\gamma}^*} = \sum_{x_i=0}^{q-1}\phi_k(x_i)\phi_l(x_i)P_i(x_i \mid \boldsymbol{\gamma}^*, \boldsymbol{J})$, the diagonal-blocks in $\boldsymbol{\chi}(\boldsymbol{m})$ are explicitly expressed in terms of $\boldsymbol{m}$ as

$$\big[\boldsymbol{\chi}(\boldsymbol{m})\big]_{e(i,k),e(i,l)} = v^{(i)}_{k,l}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]}, \tag{13}$$

where $v^{(i)}_{k,l}(\boldsymbol{m}_i) := \delta_{k,l}/q + \sum_{r=1}^{q-1} m_{[i,r]}T_{k,l,r}$ and $T_{k,l,r} := \sum_{x=0}^{q-1}\phi_k(x)\phi_l(x)\phi_r(x)$. From (12) and (13), we find that the relations

$$\big[\boldsymbol{H}(\boldsymbol{m})^{-1}\big]_{e(i,k),e(i,l)} = v^{(i)}_{k,l}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]} \tag{14}$$

are also ensured for any $\boldsymbol{m}$. Relations in (14) are referred to as *diagonal-block consistency* in this paper.

Although the diagonal-block consistency holds in the true GFE, the diagonal-block consistency can be broken if one employs an approximate scheme such as a mean-field method [12][1]. A basic concept of I-SusP is that the diagonal-block consistency is enforced in an approximate scheme, which is always kept in the exact scheme.

4. Improved Susceptibility Propagation for Pairwise MRF

In this section, an explicit formulation of I-SusP based on the Bethe approximation for (1) is proposed. First, a general scheme of I-SusP is introduced in section 4.1 in accordance with [11]. An explicit formulation of I-SusP based on a Bethe approximation is then derived in the subsequent sections.

4.1. Scheme of Improved Susceptibility Propagation

Let us suppose an approximated GFE formed by

$$\mathcal{G}_0(\boldsymbol{m}) := -\boldsymbol{h}^{\mathrm{t}}\boldsymbol{m} + \max_{\boldsymbol{\gamma}}\Big(\boldsymbol{\gamma}^{\mathrm{t}}\boldsymbol{m} + F_0(\boldsymbol{\gamma}, \boldsymbol{J})\Big), \tag{15}$$

and consider a minimization of the GFE. When $F_0(\boldsymbol{\gamma}, \boldsymbol{J})$ is the true HFE, $\mathcal{G}_0(\boldsymbol{m})$ is equal to the true GFE.

The scheme of I-SusP can be described as follows. First, with controllable parameters $\boldsymbol{\Lambda}$, we extend an approximated GFE, $\mathcal{G}_0(\boldsymbol{m})$, by adding an extra term as

$$\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda}) := \mathcal{G}_0(\boldsymbol{m}) - \frac{1}{2}\sum_{i\in V}\sum_{k,l=1}^{q-1}\Lambda_{[i,k],[i,l]}\Big(v^{(i)}_{k,l}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]}\Big). \tag{16}$$

The extra term is added so that the extended approximate GFE enables to satisfy diagonal-block consistency[2]. In the I-SusP on Ising systems, it is known that this extra term eventually behaves as an Onsager reaction term [11, 15]. The goal of I-SusP is to minimize the extended approximate GFE in (16) with respect to $\boldsymbol{m}$ with searching values of $\boldsymbol{\Lambda}$ that satisfy relations

$$\big[\boldsymbol{H}_1(\boldsymbol{m}, \boldsymbol{\Lambda})^{-1}\big]_{e(i,k),e(i,l)} = v^{(i)}_{k,l}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]}, \tag{17}$$

where $\boldsymbol{H}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$ is the Hessian matrix of $\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$ defined by

$$\big[\boldsymbol{H}_1(\boldsymbol{m}, \boldsymbol{\Lambda})\big]_{e(i,k),e(j,l)} := \frac{\partial^2\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})}{\partial m_{[i,k]}\partial m_{[j,l]}}.$$

[1] Although the investigations of the breaking in [12] addressed only the Ising case, it is expected that they can be expanded to other cases involving the multi-valued case presented in this paper.

[2] In [15], this term was introduced as the Lagrange multipliers.

Relations (17) correspond to the diagonal-block consistency in (14).

One can solve the I-SusP by using the following way. When $\boldsymbol{m} = \hat{\boldsymbol{m}}(\boldsymbol{\Lambda})$, where $\hat{\boldsymbol{m}}(\boldsymbol{\Lambda}) := \arg\min_{\boldsymbol{m}} \mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$, relations (17) can be rewritten as

$$\hat{\chi}_{[i,k],[i,l]}(\boldsymbol{\Lambda}) = v^{(i)}_{k,l}(\hat{\boldsymbol{m}}_i(\boldsymbol{\Lambda})) - \hat{m}_{[i,k]}(\boldsymbol{\Lambda})\hat{m}_{[i,l]}(\boldsymbol{\Lambda}), \tag{18}$$

where $\hat{\chi}_{[i,k],[j,l]}(\boldsymbol{\Lambda}) := \partial \hat{m}_{[i,k]}(\boldsymbol{\Lambda})/\partial h_{[j,l]}$ (see (B.2) in Appendix B). Therefore, the framework of I-SusP is summarized as follows:

STEP 1 Given $\boldsymbol{\Lambda}$, minimize $\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{m}$: $\hat{\boldsymbol{m}}(\boldsymbol{\Lambda}) = \arg\min_{\boldsymbol{m}} \mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$

STEP 2 For the given $\boldsymbol{\Lambda}$ and $\hat{\boldsymbol{m}}(\boldsymbol{\Lambda})$, compute $\hat{\chi}_{[i,k],[j,l]}(\boldsymbol{\Lambda}) = \partial \hat{m}_{[i,k]}(\boldsymbol{\Lambda})/\partial h_{[j,l]}$.

STEP 3 Solve (18) with respect to $\boldsymbol{\Lambda}$.

STEP 4 Repeat from STEP1 to STEP3 until convergence.

Equation (18) is termed *the diagonal-block matching equation* in this paper, because it leads to the matching of the diagonal-blocks of the susceptibilities obtained by an approximation with those obtained by an exact evaluation. Note that, because the parameters $\boldsymbol{\Lambda}$ are treated as independent variables in (18), the derivatives of $\boldsymbol{\Lambda}$ with respect to $\boldsymbol{h}$ are zero in the right hand side of (18). An alternative interpretation of the diagonal-block consistency in (17) and the diagonal-block equation was provided in Appendix C.

When $\mathcal{G}_0(\boldsymbol{m}) = \mathcal{G}(\boldsymbol{m})$, parameters $\boldsymbol{\Lambda}$ automatically vanish because the diagonal-block consistency always holds in the true GFE (see Appendix C). However, if $\mathcal{G}_0(\boldsymbol{m})$ is an approximate GFE that breaks the diagonal-block consistency, parameters $\boldsymbol{\Lambda}$ remain to enforce the diagonal-block consistency.

4.2. Minimization of Extended Bethe-Gibbs Free Energy

In this section, we construct the extended GFE, $\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$, by using a Bethe free energy, and subsequently, we derive a message-passing equation for minimization $\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{m}$.

In the context of the Bethe approximation based on the cluster variational method presented in [17], the HFE, $F(\boldsymbol{h}, \boldsymbol{J})$, is approximated by $F(\boldsymbol{h}, \boldsymbol{J}) \approx F_0(\boldsymbol{h}, \boldsymbol{J}) = \min_{\boldsymbol{z}} F_B(\boldsymbol{z} \mid \boldsymbol{h}, \boldsymbol{J})$, where

$$\begin{aligned} &F_B(\boldsymbol{z} \mid \boldsymbol{h}, \boldsymbol{J}) \\ &:= \min_{\boldsymbol{\xi}} \Big(- \boldsymbol{h}^{\mathrm{t}}\boldsymbol{z} - \frac{1}{2} \sum_{(i,j)\in E} \sum_{k,l=1}^{q-1} J_{[i,k],[j,l]} \xi_{[i,k],[j,l]} + \sum_{i \in V} \big(1 - |\partial(i)|\big) \sum_{x_i=0}^{q-1} \mathcal{P}_i(x_i \mid \boldsymbol{z}_i) \ln \mathcal{P}_i(x_i \mid \boldsymbol{z}_i) \\ &+ \frac{1}{2} \sum_{(i,j)\in E} \sum_{x_i,x_j=0}^{q-1} \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j) \ln \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j) \Big) \end{aligned} \tag{19}$$

is the Bethe free energy for (5) expressed by the moment representation. Marginal distributions $\mathcal{P}_i(x_i \mid \boldsymbol{z}_i)$ and $\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j)$ are defined by $\mathcal{P}_i(x_i \mid \boldsymbol{z}_i) = q^{-1} + \sum_{k=1}^{q-1} z_{[i,k]}\phi_k(x_i)$ and

$$\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j) := \frac{1}{q^2} + \frac{1}{q} \sum_{k=1}^{q-1} \big(z_{[i,k]}\phi_k(x_i) + z_{[j,k]}\phi_k(x_j) \big) + \sum_{k,l=1}^{q-1} \xi_{[i,k],[j,l]} \phi_k(x_i)\phi_l(x_j),$$

respectively [17]. It is noteworthy that the minimization of the Bethe free energy with respect to $\boldsymbol{z}$ can be done without any constraints, because, by the properties of the present orthogonal set and by the fact that $\phi_0(x)$ is a constant, marginal distributions $\mathcal{P}_i(x_i \mid \boldsymbol{z}_i)$ and $\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j)$ always satisfy the normalization and reducibility conditions, i.e., $\sum_{x_i=0}^{q-1} \mathcal{P}_i(x_i \mid \boldsymbol{z}_i) = 1$,

$\sum_{x_i,x_j=0}^{q-1} \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j) = 1$, and $\sum_{x_j=0}^{q-1} \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{z}_i, \boldsymbol{z}_j) = \mathcal{P}_i(x_i \mid \boldsymbol{z}_i)$. These can be verified by using the relation

$$\sum_{x=0}^{q-1} \phi_k(x) = \sqrt{q}\delta_{k,0}. \tag{20}$$

This is one of merits of the orthonormal expansion employed in this paper.

By employing the Bethe free energy as $F_0(\boldsymbol{\gamma}, \boldsymbol{J})$ in (15), expression $\mathcal{G}_0(\boldsymbol{m})$, namely the Bethe-Gibbs free energy, is expressed as

$$\mathcal{G}_0(\boldsymbol{m}) = -\boldsymbol{h}^{\mathrm{t}}\boldsymbol{m} + \max_{\boldsymbol{\gamma}} \big(\boldsymbol{\gamma}^{\mathrm{t}}\boldsymbol{m} + \min_{\boldsymbol{z}} F_B(\boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{J})\big).$$

The maximum condition of the right hand side of this equation with respect to $\boldsymbol{\gamma}$ leads to $\boldsymbol{m} = \boldsymbol{z}^*$, where $\boldsymbol{z}^*$ are the values of $\boldsymbol{z}$ that minimize $F_B(\boldsymbol{z} \mid \boldsymbol{\gamma}, \boldsymbol{J})$. By using this relation we can reach $\mathcal{G}_0(\boldsymbol{m}) = F_B(\boldsymbol{m} \mid \boldsymbol{h}, \boldsymbol{J})$. Therefore, the extended Bethe-Gibbs free energy corresponding to (16) is expressed as

$$\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda}) = F_B(\boldsymbol{m} \mid \boldsymbol{h}, \boldsymbol{J}) - \frac{1}{2}\sum_{i\in V}\sum_{k,l=1}^{q-1} \Lambda_{[i,k],[i,l]}\Big(v_{k,l}^{(i)}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]}\Big) \tag{21}$$

in this case.

In the following, let us derive a message-passing equation to minimize $\mathcal{G}_1(\boldsymbol{m}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{m}$. The minimum conditions of (21) with respect to $\boldsymbol{m}$ yield

$$-H_{[i,k]} + \big(1 - |\partial(i)|\big)\sum_{x_i=0}^{q-1} \phi_k(x_i)\ln \mathcal{P}_i(x_i \mid \boldsymbol{m}_i) + \frac{1}{q}\sum_{j\in\partial(i)}\sum_{x_i,x_j=0}^{q-1} \phi_k(x_i)\ln \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) = 0, \tag{22}$$

where

$$H_{[i,k]} := h_{[i,k]} + \frac{1}{2}\sum_{a,b=1}^{q-1} \Lambda_{[i,a],[i,b]}T_{a,b,k} - \sum_{a=1}^{q-1} \Lambda_{[i,k],[i,a]}m_{[i,a]}.$$

The minimum conditions in $F_B(\boldsymbol{m} \mid \boldsymbol{h}, \boldsymbol{J})$ with respect to $\boldsymbol{\xi}$ lead to

$$-J_{[i,k],[j,l]} + \sum_{x_i,x_j=0}^{q-1} \phi_k(x_i)\phi_l(x_j)\ln \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) = 0. \tag{23}$$

By combining (23) with (22), we arrive at the message-passing equation expressed by

$$\mathcal{M}_{j\to i}^{(k)} = \sum_{x_i=0}^{q-1} \phi_k(x_i)\ln \sum_{x_j=0}^{q-1} \exp\Big(\mathcal{H}_{j\to i}(x_i, x_j)\Big), \tag{24}$$

where $\mathcal{H}_{j\to i}(x_i, x_j) := \sum_{a=1}^{q-1}\big(H_{[j,a]} + \sum_{r\in\partial(j)\setminus\{i\}} \mathcal{M}_{r\to j}^{(a)}\big)\phi_a(x_j) + \sum_{a,b=1}^{q-1} J_{[i,a],[j,b]}\phi_a(x_i)\phi_b(x_j)$. The expression $\mathcal{M}_{j\to i}^{(k)}$ denotes a message from node j to node i.

By using solutions to the message-passing equation, the values of $\boldsymbol{m}$ that minimize the extended Bethe free energy (21) are expressed by

$$m_{[i,k]} = \frac{1}{Z_i} \sum_{x_i=0}^{q-1} \phi_k(x_i) \exp\Big\{ \sum_{a=1}^{q-1} \Big(H_{[i,a]} + \sum_{j\in\partial(i)} \mathcal{M}_{j\to i}^{(a)} \Big) \phi_a(x_i) \Big\}, \tag{25}$$

where expression Z_i is the normalizing constant defined by

$$Z_i := \sum_{x_i=0}^{q-1} \exp\Big\{ \sum_{a=1}^{q-1} \Big(H_{[i,a]} + \sum_{j\in\partial(i)} \mathcal{M}_{j\to i}^{(a)} \Big) \phi_a(x_i) \Big\}.$$

Equations (24) and (25) are obtained by similar manipulations presented in [17]. The details of these derivation are described in Appendix D. By solving (24) and (25), we obtain the values of $\boldsymbol{m}$ that minimize (21) for a given $\boldsymbol{\Lambda}$. Note that solutions to (24) and (25) are equivalent to those to the usual BP if $\boldsymbol{\Lambda} = \boldsymbol{0}$ are fixed.

4.3. Improved Susceptibility Propagation using Bethe Approximation

In this section, we derive the the diagonal-block equation corresponding to (18) for (21).

By differentiating (25) with respect to $h_{[j,l]}$, we obtain

$$\hat{\chi}_{[i,k],[j,l]} = \sum_{a=1}^{q-1} \Big(\delta_{i,j}\delta_{a,l} - \sum_{b=1}^{q-1} \Lambda_{[i,a],[i,b]} \hat{\chi}_{[i,b],[j,l]} + \sum_{r\in\partial(i)} \frac{\partial \mathcal{M}_{r\to i}^{(a)}}{\partial h_{[j,l]}} \Big) \big(v_{k,a}^{(i)}(\boldsymbol{m}_i) - m_{[i,k]} m_{[i,a]} \big), \tag{26}$$

where $\hat{\chi}_{[i,k],[j,l]} = \partial m_{[i,k]} / \partial h_{[j,l]}$. The derivatives of messages are obtained by solving

$$\begin{aligned} \frac{\partial \mathcal{M}_{j\to i}^{(k)}}{\partial h_{[u,v]}} = & \sum_{a=1}^{q-1} \Big(\delta_{j,u}\delta_{a,v} - \sum_{b=1}^{q-1} \Lambda_{[j,a],[j,b]} \hat{\chi}_{[j,b],[u,v]} + \sum_{r\in\partial(j)\setminus\{i\}} \frac{\partial \mathcal{M}_{r\to j}^{(a)}}{\partial h_{[u,v]}} \Big) \\ & \times \sum_{x_i=0}^{q-1} \phi_k(x_i) \frac{\sum_{x_j=0}^{q-1} \phi_a(x_j) \exp\big(\mathcal{H}_{j\to i}(x_i, x_j) \big)}{\sum_{x_j=0}^{q-1} \exp\big(\mathcal{H}_{j\to i}(x_i, x_j) \big)}. \end{aligned} \tag{27}$$

Equation (27) is obtained by differentiating (24) with respect to $h_{[u,v]}$. By letting $i = j$ in (26) and by using the relation in (18), i.e., $\hat{\chi}_{[i,k],[i,l]} = v_{k,l}^{(i)}(\boldsymbol{m}_i) - m_{[i,k]} m_{[i,l]}$, we have

$$\sum_{a,b=1}^{q-1} \Lambda_{[i,a],[i,b]} \big[\boldsymbol{\nu}_i\big]_{b,l} \big[\boldsymbol{\nu}_i\big]_{k,a} = \sum_{a=1}^{q-1} \big[\boldsymbol{\nu}_i\big]_{k,a} \sum_{j\in\partial(i)} \frac{\partial \mathcal{M}_{j\to i}^{(a)}}{\partial h_{[i,l]}}, \tag{28}$$

where $(q-1)\times(q-1)$ matrix $\boldsymbol{\nu}_i$ is defined by $\big[\boldsymbol{\nu}_i\big]_{k,l} := v_{k,l}^{(i)}(\boldsymbol{m}_i) - m_{[i,k]} m_{[i,l]}$. In the I-SusP based on the Bethe approximation, (28) corresponds to the diagonal-block matching equation in (18). Relations in (28) are collectively expressed as $\boldsymbol{\nu}_i \boldsymbol{\Lambda}_i \boldsymbol{\nu}_i = \boldsymbol{K}_i$, where $(q-1)\times(q-1)$ matrices $\boldsymbol{\Lambda}_i$ and $\boldsymbol{K}_i$ are defined by

$$\big[\boldsymbol{\Lambda}_i\big]_{k,l} := \Lambda_{\,[i,k],[i,l]} \quad \text{and} \quad \big[\boldsymbol{K}_i\big]_{k,l} := \sum_{a=1}^{q-1} \big(v_{k,a}^{(i)}(\boldsymbol{m}_i) - m_{[i,k]} m_{[i,a]} \big) \sum_{j\in\partial(i)} \frac{\partial \mathcal{M}_{j\to i}^{(a)}}{\partial h_{[i,l]}},$$

respectively. Therefore, we obtain

$$\boldsymbol{\Lambda}_i = \boldsymbol{\nu}_i^{-1} \boldsymbol{K}_i \boldsymbol{\nu}_i^{-1}. \tag{29}$$

By this equation, we can obtain the values of $\boldsymbol{\Lambda}$ that satisfy the diagonal-block matching equation in (28).

By simultaneously solving (24)–(27), and (29), we obtain solutions to the proposed I-SusP based on the Bethe approximation. It is noteworthy that the I-SusP is reduced to a usual SusP if $\boldsymbol{\Lambda} = \boldsymbol{0}$ are fixed.

5. Application to Generalized Sparse Priors for Natural Images

An image prior model using MRF, called the generalized sparse prior, which takes the form

$$\theta_i(x_i) = 0, \quad w_{(i,j)}(x_i, x_j) = -\alpha_{ij}|x_i - x_j|^p \tag{30}$$

was proposed [18], where $\alpha_{ij} = \alpha_{ji}$. This image prior includes some conventional image priors as special cases, because it is reduced to Q-Ising image priors when $p = 2$ and is reduced to Q-Potts priors when $p \to 0$. By changing the value of p, we can adjust the smoothness, the flatness, and the appearance frequency of edges in images. It is believed that values of p near 0.5 are appropriate for natural images.

In the following, let us consider a direct problem and an inverse problem on the generalized sparse prior on an $L \times L$ grid system, in which interaction parameters $\boldsymbol{\alpha}$ are independently drawn from Gaussian $p(\alpha_{ij} \mid \sigma) := (\sqrt{2\pi}\sigma^2)^{-1} \exp(-\alpha_{ij}^2/2\sigma^2)$ and the flatness parameter p is set to 0.5.

5.1. Direct Problem on Generalized Sparse Priors

Given $\boldsymbol{\alpha} = \{\alpha_{ij}\}$ and p, in order to compute statistical quantities on a generalized sparse prior, let us apply the I-SusP proposed in the previous section to the generalized sparse priors.

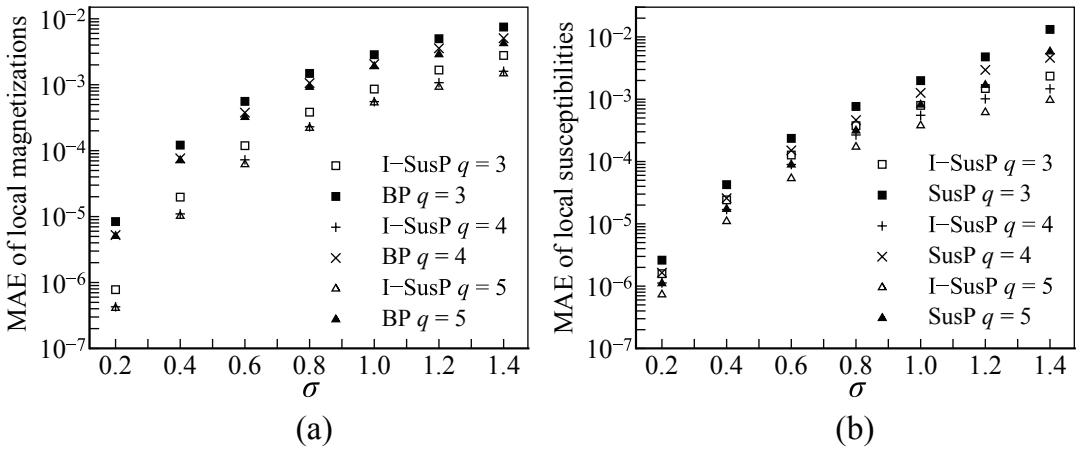

Figure 1. Plots of results of the direct problem on the generalized sparse priors when $L = 3$. (a): Comparison of local magnetizations obtained by the I-SusP with ones obtained by the BP. (b): Comparison local susceptibilities obtained by the I-SusP with those obtained by the BP. Each plot is averaged over 1000 trials.

In figure 1, mean absolute errors (MAEs) of magnetizations and susceptibilities between true values and values obtained by approximate methods against σ are plotted when $L = 3$. We see that the proposed I-SusP overcomes the normal SusP.

5.2. Inverse Problem using Observations Generated from Generalized Sparse Priors

In this section, let us solve an inverse problem, the so called inverse Ising problem, using the I-SusP without specifying the functional forms of potential functions $\{\theta_i(x_i)\}$ and $\{w_{(i,j)}(x_i, x_j)\}$. That is, we consider a non-parametric learning in this section.

Suppose that M data points, $\mathcal{D} = \{\mathbf{x}^{(\mu)} \mid \mu = 1, 2, \ldots, M\}$, are given. The aim of the inverse problem on which we focus in this section is to find the values of parameters $\boldsymbol{h}$ and $\boldsymbol{J}$ that maximize the log-likelihood function

$$l_{\mathcal{D}}(\boldsymbol{h}, \boldsymbol{J}) := \frac{1}{M} \sum_{\mu=1}^{M} \ln P_G(\mathbf{x}^{(\mu)} \mid \boldsymbol{h}, \boldsymbol{J}). \tag{31}$$

The derivatives of the log-likelihood function with respect to $\boldsymbol{h}$ and $\boldsymbol{J}$ are

$$\Delta_{h_{[i,k]}}(\boldsymbol{h}, \boldsymbol{J}) := \frac{\partial l_{\mathcal{D}}(\boldsymbol{h}, \boldsymbol{J})}{\partial h_{[i,k]}} = \langle \phi_k(x_i) \rangle_{\mathcal{D}} - \sum_{\boldsymbol{x}} \phi_k(x_i) P_G(\boldsymbol{x} \mid \boldsymbol{h}, \boldsymbol{J})$$

and

$$\Delta_{J_{[i,k],[j,l]}}(\boldsymbol{h}, \boldsymbol{J}) := \frac{\partial l_{\mathcal{D}}(\boldsymbol{h}, \boldsymbol{J})}{\partial J_{[i,k],[j,l]}} = \langle \phi_k(x_i)\phi_l(x_j) \rangle_{\mathcal{D}} - \sum_{\boldsymbol{x}} \phi_k(x_i)\phi_l(x_j) P_G(\boldsymbol{x} \mid \boldsymbol{h}, \boldsymbol{J}),$$

respectively, where $\langle \cdots \rangle_{\mathcal{D}}$ is a sample average of given data points. Because the log-likelihood function (31) is a concave function with respect to the parameters, a maximum point of the log-likelihood can be found by a gradient ascent method:

$$h^{\text{new}}_{[i,k]} = h^{\text{old}}_{[i,k]} + \eta \Delta_{h_{[i,k]}}(\boldsymbol{h}^{\text{old}}, \boldsymbol{J}^{\text{old}}), \tag{32}$$

$$J^{\text{new}}_{[i,k],[j,l]} = J^{\text{old}}_{[i,k],[j,l]} + \eta \Delta_{J_{[i,k],[j,l]}}(\boldsymbol{h}^{\text{old}}, \boldsymbol{J}^{\text{old}}), \tag{33}$$

where η is a small positive constant.

To apply the I-SusP to the inverse problem, the derivatives are approximated by

$$\Delta_{h_{[i,k]}}(\boldsymbol{h}, \boldsymbol{J}) \approx \Delta^{\text{app}}_{h_{[i,k]}}(\boldsymbol{h}, \boldsymbol{J}) := \langle \phi_k(x_i) \rangle_{\mathcal{D}} - m_{i,k}, \tag{34}$$

$$\Delta_{J_{[i,k],[j,l]}}(\boldsymbol{h}, \boldsymbol{J}) \approx \Delta^{\text{app}}_{J_{[i,k],[j,l]}}(\boldsymbol{h}, \boldsymbol{J}) := \langle \phi_k(x_i)\phi_l(x_j) \rangle_{\mathcal{D}} - \left(\bar{\chi}_{[i,k],[j,l]} + m_{i,k}m_{j,l}\right), \tag{35}$$

where $\boldsymbol{m}$ and $\bar{\boldsymbol{\chi}}$ are solutions to the I-SusP. By solving (24)–(27), (29), and (32) and (33) with the approximations in (34) and (35), we can solve the inverse problem by using the I-SusP for the given data set $\mathcal{D}$. If $\boldsymbol{\Lambda} = \mathbf{0}$ are fixed, we can obtain solutions based on the SusP. In the following, results of the inverse problem using data points generated from the generalized sparse priors are shown.

In figure 2, MAEs between true biases $\boldsymbol{h}$ and interactions $\boldsymbol{J}$ and those obtained by approximate methods are plotted for $q = 3$ and $q = 4$ when $L = 3$. Here, instead of sampled expectations, true statistical values of the generalized sparse priors are used, namely, we consider an identical case where $M \to \infty$. The results of the BP in these plots are obtained by using a non-iterative method proposed in [17]. The results obtained by the SusP are poor and nearly the same as those obtained by the BP in the region of large σ. In contrast, the results obtained by the I-SusP are improved for the whole region.

Results of the inverse problem when $L = 8$ and $q = 3$ for $\sigma = 0.4$ and $\sigma = 0.8$ are shown in table 1. In these experiments, $M = 10000$ data points, that are generated by using the Markov chain Monte Carlo (MCMC) method on the generalized sparse priors, are used. We can see that the I-SusP gives the best results.

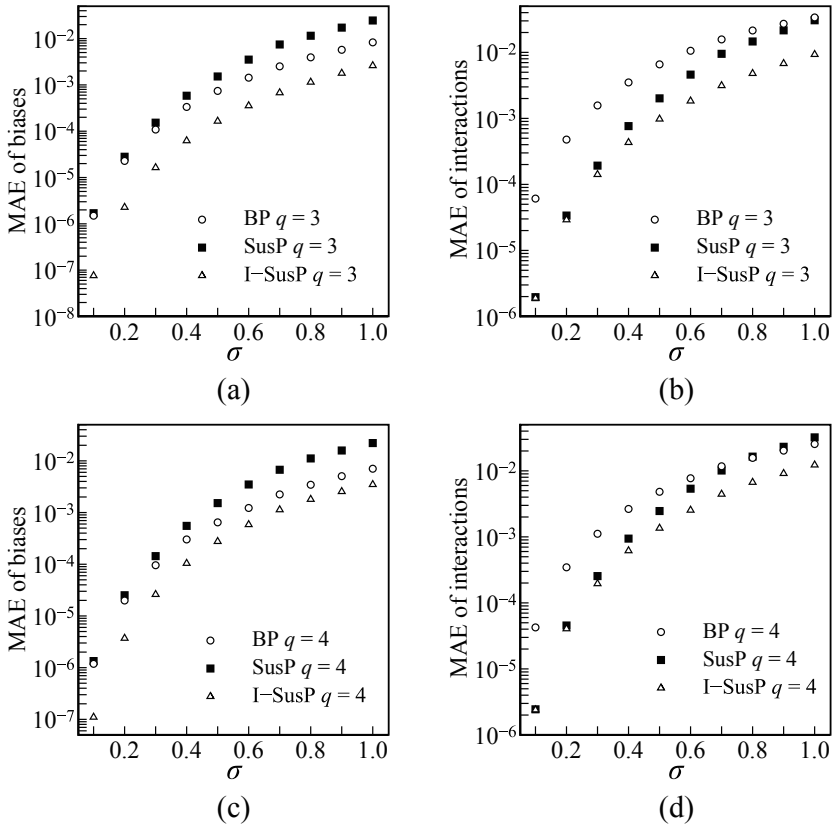

Figure 2. Plots of results of the inverse problem on the generalized sparse priors when $L = 3$. (a) and (c): MAEs of biases obtained by the BP, the SusP, and the I-SusP when $q = 3$ and $q = 4$, respectively. (b) and (d): MAEs of interactions obtained by the BP, the SusP, and the I-SusP when $q = 3$ and $q = 4$, respectively. Each plot is averaged over 1000 trials.

Table 1. Results of the inverse problem on the generalized sparse priors when $L = 8$ and $q = 3$. $M = 10000$ data points generated by using MCMC method on the generalized sparse prior are used. Each value is averaged over 30 trials.

	σ	MAE of biases	MAE of interactions
BP	0.4	0.0153	0.0262
	0.8	0.0197	0.0417
SusP	0.4	0.0153	0.0248
	0.8	0.0295	0.0365
I-SusP	0.4	0.0152	0.0246
	0.8	0.0184	0.0278

6. Conclusion and Discussion

In this paper, the generalization scheme of I-SusP using the orthogonal expansion was proposed in which any pairwise potential functions and multi-valued random variables are acceptable. Moreover, the proposed I-SusP is applied to the direct and the inverse problems on generalized sparse priors for natural images. The constraints of diagonal-block consistencies improved the

performances of the algorithms and yielded results that surpass those obtained by conventional methods.

Although the orthogonal expansion using Chebyshev polynomials employed in this paper should not be the only possible choice for a generalization, it enables us to reach moment representations of free energies as presented in (19) and can make subsequent manipulations clear. This could help us in future developments, especially in inverse problems, as it helped us in inverse problems of the normal BP level [17].

In the usual SusP scheme, we first solve the BP ((24) and (25) with $\mathbf{\Lambda} = \mathbf{0}$), and obtain solutions to the BP. These solutions are the values of the local magnetizations and the messages. Subsequently, to obtain the susceptibilities, we solve the SusP ((26) and (27) with $\mathbf{\Lambda} = \mathbf{0}$) by using the solutions of the BP. Therefore, in the conventional scheme shown in figure 3 (a), the results of the SusP have no effect on the BP. In contrast, the scheme of I-SusP shown in

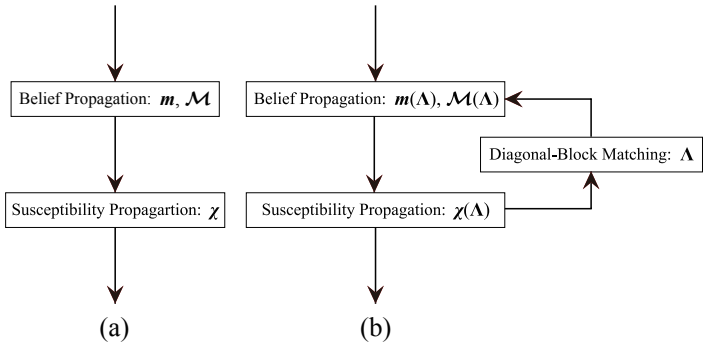

Figure 3. Illustration of schemes of (a) normal SusP and (b) I-SusP.

figure 3 (b) includes feedback from the SusP to the BP through parameters $\mathbf{\Lambda}$. This feedback marks a significant difference between the scheme of conventional SusP and the scheme of I-SusP. It is noteworthy that the computational cost of solving the I-SusP is the same as that of solving the usual SusP, because the diagonal-block matching equation in (29) does not contribute to the order of the total computational cost.

One of the most interesting future applications is a direct and an inverse problem in the deep Boltzmann machine [19, 20], which is a Boltzmann machine with a hierarchical structure and one of the most important learning models in the recent machine learning field.

Acknowledgments
This work was partly supported by Grants-In-Aid (No. 24700220) for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Appendix A. Chebyshev Polynomials

One possible choice of orthonormal set is given by [21]

$$\phi_k(x) := (-1)^k \psi_k(x) \sqrt{\frac{(2k+1)[(q-1)!]^2}{(q+k)!(q-k-1)!}} \tag{A.1}$$

for $x = 0, 1, \ldots, q-1$ and $k = 0, 1, \ldots, q-1$, where $\{\psi_k(x) \mid k = 0, 1, \ldots, q-1\}$ are the Chebyshev polynomials in the discrete range given by

$$\psi_k(x) := \sum_{l=0}^{k} (-1)^l \binom{k}{l} \binom{k+l}{l} \frac{x!(q-l-1)!}{\Gamma(x-l+1)(q-1)!}, \tag{A.2}$$

where $\Gamma(x)$ is the gamma function. The Chebyshev polynomials (A.2) are also given by the recursion formula

$$(k+1)(q-1-k)\psi_{k+1}(x) = -(2x-q+1)(2k+1)\psi_k(x) - k(q+k)\psi_{k-1}(x),$$

starting from $\psi_0(x) = 1$ and $\psi_1(x) = 1 - 2x/(q-1)$. From (A.1) and (A.2), we have

$$\phi_0(x) = 1/\sqrt{q}, \phi \quad _1(x) = \sqrt{3}(2x-q+1)/\sqrt{q(q^2-1)},$$

and so on.

Appendix B. Linear Response Relation for Approximate GFEs

For $\boldsymbol{m} = \{m_{[i,k]} \mid i \in V; k = 1, 2, \ldots, q-1\}$, let us consider a function expressed in the form

$$\hat{\mathcal{G}}(\boldsymbol{m}) = -\boldsymbol{h}^{\mathrm{t}}\boldsymbol{m} + \hat{g}(\boldsymbol{m}).$$

It is to be noted that this expression includes all the GFEs ($\mathcal{G}(\boldsymbol{m})$, $\mathcal{G}_0(\boldsymbol{m})$, and $\mathcal{G}_1(\boldsymbol{m})$) presented in this paper. At a minimum of the function, from its extremal condition, relations

$$-h_{[i,k]} + \left.\frac{\partial \hat{g}(\boldsymbol{m})}{\partial m_{[i,k]}}\right|_{\boldsymbol{m}=\boldsymbol{m}^\star} = 0 \tag{B.1}$$

hold, where $\boldsymbol{m}^\star := \arg\min_{\boldsymbol{m}} \hat{\mathcal{G}}(\boldsymbol{m})$. By differentiating (B.1) with respect to $h_{[j,l]}$, we obtain

$$\delta_{i,j} = \sum_{r\in V}\sum_{a=1}^{q-1} \left.\frac{\partial^2 \hat{g}(\boldsymbol{m})}{\partial m_{[i,k]}\partial m_{[r,a]}}\right|_{\boldsymbol{m}=\boldsymbol{m}^\star} \frac{\partial m^\star_{[r,a]}}{\partial h_{[j,l]}} = \sum_{r\in V}\sum_{a=1}^{q-1} \left.\frac{\partial^2 \hat{\mathcal{G}}(\boldsymbol{m})}{\partial m_{[i,k]}\partial m_{[r,a]}}\right|_{\boldsymbol{m}=\boldsymbol{m}^\star} \frac{\partial m^\star_{[r,a]}}{\partial h_{[j,l]}}.$$

This expression indicates that the relation

$$\hat{\boldsymbol{H}}^{-1} = \boldsymbol{\chi}^\star \tag{B.2}$$

holds, where matrix $\hat{\boldsymbol{H}}$ denotes the Hessian matrix of $\hat{\mathcal{G}}(\boldsymbol{m})$ defined by

$$\left[\hat{\boldsymbol{H}}(\boldsymbol{m})\right]_{e(i,k),e(j,l)} := \left.\frac{\partial^2 \hat{\mathcal{G}}(\boldsymbol{m})}{\partial m_{[i,k]}\partial m_{[j,l]}}\right|_{\boldsymbol{m}=\boldsymbol{m}^\star},$$

and matrix $\boldsymbol{\chi}^\star$ is defined by $[\boldsymbol{\chi}^\star]_{e(i,k),e(j,l)} := \partial m^\star_{[i,k]}/\partial h_{[j,l]}$.

Appendix C. Variational Interpretation of Diagonal-Block Equation

The diagonal-block matching equation in (18) is introduced to overcome the diagonal-block inconsistency problem. Here, we provide an alternative interpretation of the equation.

First, we define a measure of closeness between two positive definite symmetric matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$, in terms of the Kullback-Leibler divergence as

$$D(\boldsymbol{A}||\boldsymbol{B}) := \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{x}\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{A}) \ln \frac{\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{A})}{\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{B})},$$

where $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\Sigma}) \propto \exp(-2^{-1}\boldsymbol{x}^{\mathrm{t}}\boldsymbol{\Sigma}\boldsymbol{x})$ represents the multivariate zero-mean Gaussian with the covariance matrix $\boldsymbol{\Sigma}^{-1}$. From the properties of the Kullback-Leibler divergence, $D(\boldsymbol{A}||\boldsymbol{B}) \geq 0$ and $D(\boldsymbol{A}||\boldsymbol{B}) = 0$ if and only if $\boldsymbol{A} = \boldsymbol{B}$ is ensured. Let us consider the quantity

$D(\boldsymbol{H}(\boldsymbol{m})||\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda}))$, where $\boldsymbol{H}(\boldsymbol{m})$ is the Hessian matrix of true GFE defined in section 3.2 and $\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda})$ is the Hessian matrix of $\mathcal{G}_1(\boldsymbol{m},\boldsymbol{\Lambda})$ defined in (4.1). Minimizing $D(\boldsymbol{H}(\boldsymbol{m})||\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda}))$ corresponds to minimizing the distance [3] between the true Hessian matrix and Hessian matrix $\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda})$.

The minimum conditions of $D(\boldsymbol{H}(\boldsymbol{m})||\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda}))$ with respect to $\boldsymbol{\Lambda}$ are

$$v^{(i)}_{k,l}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]} = \left[\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda})^{-1}\right]_{e(i,k),e(i,l)}, \tag{C.1}$$

where we use the fact that $\left[\boldsymbol{H}(\boldsymbol{m},\boldsymbol{\Lambda})^{-1}\right]_{e(i,k),e(i,l)} = v^{(i)}_{k,l}(\boldsymbol{m}_i) - m_{[i,k]}m_{[i,l]}$ holds for any $\boldsymbol{m}$ in (14). The minimum conditions in (C.1) correspond to (17) for any $\boldsymbol{m}$. When $\boldsymbol{m} = \hat{\boldsymbol{m}}(\boldsymbol{\Lambda})$, upon using (B.2), (C.1) yields (18). Therefore, we can reinterpret the diagonal-block matching equation as the condition of minimization of distance between the true Hessian matrix and its approximation in terms of the Kullback-Leibler divergence at $\boldsymbol{m} = \hat{\boldsymbol{m}}(\boldsymbol{\Lambda})$.

When $\mathcal{G}_0(\boldsymbol{m}) = \mathcal{G}(\boldsymbol{m})$, minimization of $D(\boldsymbol{H}(\boldsymbol{m})||\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda}))$ obviously leads to $\boldsymbol{\Lambda} = \boldsymbol{0}$ because $\boldsymbol{H}_1(\boldsymbol{m},\boldsymbol{\Lambda}) = \boldsymbol{H}(\boldsymbol{m},\boldsymbol{\Lambda}) + \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the diagonal-block matrix defined by $[\boldsymbol{\Lambda}]_{e(i,k),e(j,l)} := \delta_{i,j}\Lambda_{[i,k],[i,l]}$.

Appendix D. Derivation of Message-Passing Equation in (24)

As mentioned in section 4.2, the one-variable and the two-variable marginal distributions are expressed as

$$\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) = \frac{1}{q} + \sum_{k=1}^{q-1} m_{[i,k]}\phi_k(x_i), \tag{D.1}$$

$$\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) = \frac{1}{q^2} + \frac{1}{q}\sum_{k=1}^{q-1}\left(m_{[i,k]}\phi_k(x_i) + m_{[j,k]}\phi_k(x_j)\right) + \sum_{k,l=1}^{q-1}\xi_{[i,k],[j,l]}\phi_k(x_i)\phi_l(x_j). \tag{D.2}$$

Since relations $\sum_{x_j=0}^{q-1}\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) = \mathcal{P}_i(x_i \mid \boldsymbol{m}_i)$ always hold, relations

$$\sum_{x_i=0}^{q-1}\phi_k(x_i)\ln\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) = \sum_{x_i=0}^{q-1}\phi_k(x_i)\ln\sum_{x_j=0}^{q-1}\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) \tag{D.3}$$

also hold.

We can express (22) as

$$-H_{[i,k]} + \sum_{x_i=0}^{q-1}\phi_k(x_i)\ln\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) - \sum_{j\in\partial(i)}\mathcal{M}^{(k)}_{j\to i} = 0, \tag{D.4}$$

where

$$\mathcal{M}^{(k)}_{j\to i} := \sum_{x_i=0}^{q-1}\phi_k(x_i)\ln\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) - \frac{1}{q}\sum_{x_i,x_j=0}^{q-1}\phi_k(x_i)\ln\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j). \tag{D.5}$$

[3] This measure is generally not a distance in a precise mathematical sense, because $D(\boldsymbol{A}||\boldsymbol{B}) \neq D(\boldsymbol{B}||\boldsymbol{A})$.

Without loss of generalities, we can rewrite $\mathcal{P}_i(x_i \mid \boldsymbol{m}_i)$ as

$$\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) = \exp\big(\ln \mathcal{P}_i(x_i \mid \boldsymbol{m}_i)\big) \propto \exp\Big(\sum_{k=1}^{q-1} c_{[i,k]}\phi_k(x_i)\Big), \tag{D.6}$$

where $c_{[i,k]} := \sum_{x_i=0}^{Q-1} \phi_k(x_i) \ln \mathcal{P}_i(x_i \mid \boldsymbol{m}_i)$ and relation $\phi_0(x_i) = 1/\sqrt{q}$ is used. Substituting (D.6) in (D.4) and using (20), we get

$$c_{[i,k]} = H_{[i,k]} + \sum_{j\in\partial(i)} \mathcal{M}_{j\to i}^{(k)}. \tag{D.7}$$

Therefore, from (D.6) and (D.7), $\mathcal{P}_i(x_i \mid \boldsymbol{m}_i)$ is expressed as

$$\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) = \frac{1}{Z_i} \exp\Big\{ \sum_{k=1}^{Q-1} \Big(H_{[i,k]} + \sum_{j\in\partial(i)} \mathcal{M}_{j\to i}^{(k)}\Big)\phi_k(x_i)\Big\}, \tag{D.8}$$

where Z_i is the normalization constant. On the other hand, from (D.1) and (20), relations

$$\sum_{x_i=0}^{q-1} \phi_k(x_i)\mathcal{P}_i(x_i \mid \boldsymbol{m}_i) = m_{[i,k]} \tag{D.9}$$

should always hold. Combining (D.8) with (D.9), we arrive at (25).

In the following, we derive an alternative expression of the two-variable marginal distribution in a similar way to above derivation. Without loss of generalities, we can rewrite $\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j)$ as

$$\begin{aligned}\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) &= \exp\big(\ln \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j)\big) \\ &\propto \exp\Big(\frac{1}{\sqrt{q}} \sum_{k=1}^{q-1} d_{[i,k],[j,0]}\phi_k(x_i) + \frac{1}{\sqrt{q}} \sum_{l=1}^{q-1} d_{[i,0],[j,l]}\phi_l(x_j) + \sum_{k,l=1}^{q-1} d_{[i,k],[j,l]}\phi_k(x_i)\phi_l(x_j)\Big),\end{aligned} \tag{D.10}$$

where $d_{[i,k],[j,l]} := \sum_{x_i,x_j=0}^{q-1} \phi_k(x_i)\phi_l(x_j) \ln \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j)$. From (23), we find

$$d_{[i,k],[j,l]} = J_{[i,k],[j,l]} \tag{D.11}$$

for $1 \le k \le q-1$ and $1 \le l \le q-1$. Using (D.4), (D.5), and the definition of $d_{[i,k],[j,l]}$, we get

$$\begin{aligned}\mathcal{M}_{j\to i}^{(k)} &= \sum_{x_i=0}^{q-1} \phi_k(x_i) \ln \mathcal{P}_i(x_i \mid \boldsymbol{m}_i) - \frac{1}{\sqrt{q}} \sum_{x_i,x_j=0}^{q-1} \phi_k(x_i)\phi_0(x_j) \ln \mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) \\ &= H_{[i,k]} + \sum_{j\in\partial(i)} \mathcal{M}_{j\to i}^{(k)} - \frac{1}{\sqrt{q}} d_{[i,k],[j,0]}.\end{aligned} \tag{D.12}$$

Similarly, we can obtain

$$\mathcal{M}_{i\to j}^{(l)} = H_{[j,l]} + \sum_{i\in\partial(j)} \mathcal{M}_{i\to j}^{(l)} - \frac{1}{\sqrt{q}} d_{[i,0],[j,l]}. \tag{D.13}$$

From (D.10)–(D.13), we find

$$\mathcal{P}_{(i,j)}(x_i, x_j \mid \boldsymbol{m}_i, \boldsymbol{m}_j) = \frac{1}{Z_{(i,j)}} \exp\Big\{ \sum_{k=1}^{q-1} \Big(H_{[i,k]} + \sum_{r\in\partial(i)\backslash\{j\}} \mathcal{M}^{(k)}_{r\to i} \Big) \phi_k(x_i) + \sum_{l=1}^{q-1} \Big(H_{[j,l]} + \sum_{r\in\partial(j)\backslash\{i\}} \mathcal{M}^{(l)}_{r\to j} \Big) \phi_l(x_j) + \sum_{k=1}^{Q-1} \sum_{l=l}^{Q-1} J^{(k,l)}_{(i,j)} \phi_k(x_i)\phi_l(x_j) \Big\}, \tag{D.14}$$

where $Z_{(i,j)}$ is the normalization constant.

Substituting (D.8) and (D.14) in (D.3) and using (2) and (20), we can reach the message-passing equation in (24).

References

[1] Opper M and Saad D (eds) 2001 *Advanced Mean Field Methods—Theory and Practice* (MIT Press)
[2] Mézard M and Montanari A 2009 *Information, Physics and Computation* (Oxford University Press)
[3] Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd ed.)* (San Francisco, CA: Morgan Kaufmann)
[4] Bethe H A 1935 *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* **150** 552–75
[5] Kabashima Y and Saad D 1998 *Europhys. Lett.* **44** 668–74
[6] Yedidia J S, Freeman W T and Weiss Y 2001 *Advances in Neural Information Processing Systems* vol 13 pp 689–95
[7] Tanaka K 2003 *IEICE Trans. on Information and Systems* **E86-D** 1228–42
[8] Welling M and Teh Y W 2003 *Artifcial Intelligence* **143** 19–50
[9] Welling M and Teh Y W 2004 *Neural Computation* **16** 197–221
[10] Mézard M and Mora T 2009 *Journal of Physiology-Paris* **103** 107–13
[11] Yasuda M and Tanaka K 2013 *Phys. Rev.* E **87** 012134
[12] Yasuda M and Tanaka K 2007 *J. Phys. A: Math. and Theor.* **40** 9993–10007
[13] Opper M and Winther O 2001 *Phys. Rev. Lett.* **86** 3695–9
[14] Opper M and Winther O 2001 *Phys. Rev.* E **64** 056131
[15] Raymond J and Ricci-Tersenghi F 2013 *In IEEE ICC'13 - Workshop on Networking across disciplines: Communication Networks, Complex Systems and Statistical Physics (NETSTAT) (ICC'13 - IEEE ICC'13 Workshop NETSTAT)*
[16] Raymond J and Ricci-Tersenghi F 2013 *Phys. Rev.* E **87** 05211
[17] Yasuda M, Kataoka S and Tanaka K 2012 *J. Phys. Soc. Jpn.* **81** 044801
[18] Tanaka K, Yasuda M and Titterington D M 2012 *J. Phys. Soc. Jpn.* **81** 114802
[19] Salakhutdinov R and Hinton G E 2009 *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)* (Clearwater Beach, FL) pp 448–55
[20] Salakhutdinov R and Hinton G E 2012 *Neural Computation* **24** 1967–2006
[21] Morita T 1993 *J. Phys. Soc. Jpn.* **62** 4218–23