Relationship between Pretraining and Maximum Likelihood Estimation in Deep Boltzmann Machines

Muneki Yasuda
Graduate School of Science and Engineering, Yamagata University

Abstract

A pretraining algorithm, which is a layer-by-layer greedy learning algorithm, for a deep Boltzmann machine (DBM) is presented in this paper. By considering the deep belief net type of pretraining for the DBM, which is a simplified version of the original pretraining of the DBM, two interesting theoretical facts about pretraining can be obtained. (1) By applying two different types of approximation, a replacing approximation by using a Bayesian network and a Bethe type of approximation based on the cluster variation method, to two different parts of the true log-likelihood function of the DBM, pretraining can be derived from a variational approximation of the original maximum likelihood estimation. (2) It can be ensured that the pretraining improves the variational bound of the true log-likelihood function of the DBM. These two theoretical results will help deepen our understanding of deep learning. Moreover, on the basis of the theoretical results, we discuss the original pretraining of the DBM in the latter part of this paper.

1 Introduction

Pretraining is one of the central techniques for training learning models with deep architectures in which we perform greedy layer-wise unsupervised training from the bottom to the top (Hinton et al., 2006). Although pretraining seems to be rather heuristic and far from principle learning strategies for models, e.g., the backpropagation method for deep neural networks, it is known in practice that they can produce powerful models that outperform models obtained by conventional techniques, such as a support vector machine (Bengio, 2009). It is believed that models trained by pretraining are certain stacked autoencoders and learn high-level representations of observed data sets, which is why pretraining is effective (Hinton and Salakhutdinov, 2006; Bengio, 2009). Empirical knowledge about pretraining has been accumulated in this decade. However, theoretical knowledge, such as the relationships between the principle learning methods and pretraining, is lacking. It is believed that the development of a theoretical background for pretraining will increase our understanding of deep learning.

A deep Boltzmann machine (DBM) (Salakhutdinov and Hinton, 2009) is a probabilistic deep learning model proposed as an extension of a deep belief net (DBN) (Hinton et al., 2006). Although, in principle, DBMs should be trained by maximum likelihood estimation (MLE), pretraining is applied to them, and the resulting DBMs provide an excellent performance in various applications, such as pattern recognition systems. In the pretraining of DBNs, each of the two layers is regarded as a restricted Boltzmann machine (RBM), which is separated from the other layers, and an unsupervised learning algorithm, such as contrastive divergence (Hinton, 2002) or persistent contrastive divergence (Tieleman, 2008), is run for each RBM separately. Although the procedure of pretraining for DBMs is almost the same as that for DBNs, it is slightly modified by, e.g., replicating the bottom and top layers and doubling the weights in the intermediate layers (Salakhutdinov and Hinton, 2009; Salakhutdinov and Larochelle, 2010; Salakhutdinov and Hinton, 2012).

The aim of this paper is to provide some insight into the relationship between pretraining and MLE in a DBM. In the first part, a simplified version of pretraining for a DBM without modification, i.e., the DBN type of pretraining for the DBM, is considered, and two interesting theoretical facts about pretraining procedure are revealed: (1) it is derived from a variational approximation of the MLE procedure, and (2) it

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

improves the variational lower bound of the true log-likelihood function of the DBM. In the latter part, we apply these two theoretical results to the original type of pretraining in DBMs and discuss it. A variational analysis of DBMs was presented by Salakhutdinov and Hinton (Salakhutdinov and Hinton, 2012). However, their analysis is applicable to only a specific type of DBM. In contrast, the analysis presented in this paper is applicable to general DBMs.

2 Deep Boltzmann Machine

Let us consider a DBM with R hidden layers and denote the set of visible variables and the set of hidden variables in the rth hidden layer by $\boldsymbol{v} := \{v_i \mid i \in \mathcal{V}\}$ and $\boldsymbol{h}^{(r)} := \{h_j^{(r)} \mid j \in \mathcal{H}_r\}$, respectively, where $\mathcal{V}$ and $\mathcal{H}_r$ are the sets of labels of variables in the visible layer and rth hidden layer, respectively. The visible and the hidden variables are assumed to be discrete. The energy function of the DBM is expressed by

$$E_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H}; \boldsymbol{\theta}) := -\sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{H}_1} w_{ij}^{(1)} v_i h_j^{(1)} - \sum_{r=2}^{R} \sum_{i \in \mathcal{H}_{r-1}} \sum_{j \in \mathcal{H}_r} w_{ij}^{(r)} h_i^{(r-1)} h_j^{(r)}, \quad (1)$$

where $\boldsymbol{\theta} := \{\boldsymbol{w}^{(r)} \mid r = 1, 2, \ldots, R\}$ is the set of the connection parameters between two layers and $\boldsymbol{H} := \{\boldsymbol{h}^{(r)} \mid r = 1, 2, \ldots, R\}$. The DBM is the probabilistic deep learning model described by

$$P_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H} \mid \boldsymbol{\theta}) := \frac{1}{Z_{\mathrm{DBM}}(\boldsymbol{\theta})} \exp\big(-E_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H}; \boldsymbol{\theta})\big), \quad (2)$$

where $Z_{\mathrm{DBM}}(\boldsymbol{\theta}) := \sum_{\boldsymbol{v}} \sum_{\boldsymbol{H}} \exp(-E_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H}; \boldsymbol{\theta}))$ is the partition function of the DBM. In Fig. 1, an example of a DBM with three hidden layers (Fig. 1(a)) and its simple representation (Fig. 1(b)) are shown.

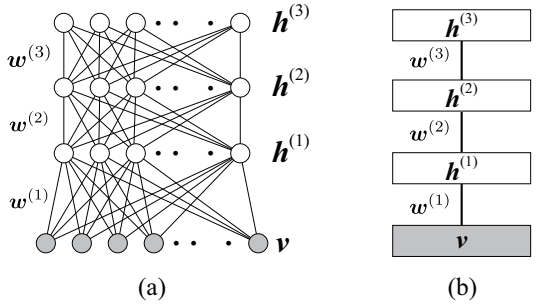

Figure 1: (a) DBM with three hidden layers ($R = 3$). (b) Simple representation of (a).

For a given set of N data points $\mathcal{D} := \{\mathbf{v}^{(\mu)} \mid \mu = 1, 2, \ldots, N\}$, the log-likelihood function of the DBM is described as

$$L_{\mathrm{DBM}}(\boldsymbol{\theta}) := \sum_{\boldsymbol{v}} Q_{\mathcal{D}}(\boldsymbol{v}) \ln P_{\mathrm{DBM}}(\boldsymbol{v} \mid \boldsymbol{\theta}), \quad (3)$$

where $P_{\mathrm{DBM}}(\boldsymbol{v} \mid \boldsymbol{\theta})$ is the marginal distribution of Eq. (2), and

$$Q_{\mathcal{D}}(\boldsymbol{v}) := \frac{1}{N} \sum_{\mu=1}^{N} \delta(\boldsymbol{v}, \mathbf{v}^{(\mu)})$$

is the empirical distribution of the data points. Here, $\delta(\boldsymbol{a}, \boldsymbol{b})$ is the Kronecker delta. By maximizing Eq. (3) with respect to $\boldsymbol{\theta}$, we achieve the maximum likelihood estimate of the DBM. The MLE procedure of the DBM, however, is computationally expensive; therefore, in practice, we train the DBM by using *layer-wise pretraining* to find good initial values of $\boldsymbol{\theta}$, and after pretraining, we fine-tune the parameters using an approximate MLE procedure such as the mean-field approximation.

3 DBN Type of Pretraining for DBMs

Although the original pretraining procedure for DBMs is not the same as that for DBNs (Salakhutdinov and Hinton, 2009; Salakhutdinov and Larochelle, 2010; Salakhutdinov and Hinton, 2012), we first consider the DBN type of pretraining (Hinton et al., 2006) for the DBM. We detail the pretraining procedure as follows. First, let us define the layer-wise RBM. For $1 \leq r \leq R$, the RBM consisting of the $(r-1)$th and rth hidden layers is expressed as

$$P_{\mathrm{RBM}}^{(r)}(\boldsymbol{h}^{(r-1)}, \boldsymbol{h}^{(r)} \mid \boldsymbol{w}^{(r)}) := \frac{1}{Z_{\mathrm{RBM}}^{(r)}(\boldsymbol{w}^{(r)})} \exp\big(-E_{\mathrm{RBM}}^{(r)}(\boldsymbol{h}^{(r-1)}, \boldsymbol{h}^{(r)}; \boldsymbol{w}^{(r)})\big), \quad (4)$$

where $E_{\mathrm{RBM}}^{(r)}(\boldsymbol{h}^{(r-1)}, \boldsymbol{h}^{(r)}; \boldsymbol{w}^{(r)})$ is the energy function of the RBM defined by

$$E_{\mathrm{RBM}}^{(r)}(\boldsymbol{h}^{(r-1)}, \boldsymbol{h}^{(r)}; \boldsymbol{w}^{(r)}) := -\sum_{i \in \mathcal{H}_{r-1}} \sum_{j \in \mathcal{H}_r} w_{ij}^{(r)} h_i^{(r-1)} h_j^{(r)},$$

and $Z_{\mathrm{RBM}}^{(r)}(\boldsymbol{w}^{(r)})$ is the partition function of the RBM. Here and in the following, the zeroth hidden layer is identified as the visible layer, i.e., $\boldsymbol{h}^{(0)} = \boldsymbol{v}$ and $\mathcal{H}_0 = \mathcal{V}$.

During pretraining, we first train $\boldsymbol{w}^{(1)}$ by training the first RBM, $P_{\mathrm{RBM}}^{(1)}(\boldsymbol{v}, \boldsymbol{h}^{(1)} \mid \boldsymbol{w}^{(1)})$, consisting of the visible layer and first hidden layer (RBM1 in Fig. 2(a))

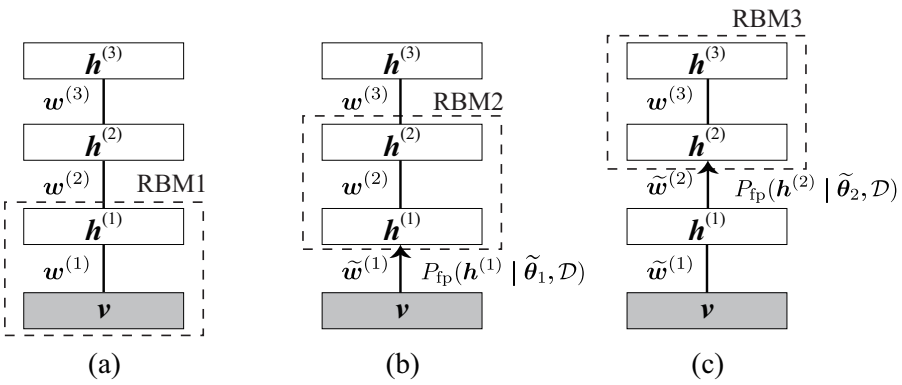

Figure 2: DBN type of pretraining for the DBM in Fig. 1.

using MLE with the given data set $\mathcal{D}$, namely, maximizing the log-likelihood function

$$L^{(1)}_{\mathrm{RBM}}(\boldsymbol{w}^{(1)}) := \sum_{\boldsymbol{v}} Q_{\mathcal{D}}(\boldsymbol{v}) \ln \sum_{\boldsymbol{h}^{(1)}} P^{(1)}_{\mathrm{RBM}}(\boldsymbol{v}, \boldsymbol{h}^{(1)} \mid \boldsymbol{w}^{(1)}) \tag{5}$$

with respect to $\boldsymbol{w}^{(1)}$. Let the solution for training the first RBM be represented by $\tilde{\boldsymbol{w}}^{(1)}$.

After training the first RBM, we train $\boldsymbol{w}^{(2)}$ using the maximum likelihood estimate of the second RBM, $P^{(2)}_{\mathrm{RBM}}(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \boldsymbol{w}^{(2)})$, consisting of the first and second hidden layers (RBM2 in Fig. 2(b)). During training, we use *the feature points* in the first hidden layer, which are generated by

$$P_{\mathrm{fp}}(\boldsymbol{h}^{(1)} \mid \tilde{\boldsymbol{w}}^{(1)}, \mathcal{D}) := \sum_{\boldsymbol{v}} P^{(1)}_{\mathrm{RBM}}(\boldsymbol{h}^{(1)} \mid \boldsymbol{v}, \tilde{\boldsymbol{w}}^{(1)}) Q_{\mathcal{D}}(\boldsymbol{v}),$$

where $P^{(r)}_{\mathrm{RBM}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{h}^{(r-1)}, \boldsymbol{w}^{(r)})$ is the conditional distribution of the RBM in Eq. (4) from the lower layer to the upper layer, as the pseudo-data points. Namely, when training the second RBM, we maximize the log-likelihood function

$$L^{(2)}_{\mathrm{RBM}}(\boldsymbol{w}^{(2)}; \tilde{\boldsymbol{w}}^{(1)}) := \sum_{\boldsymbol{h}^{(1)}} P_{\mathrm{fp}}(\boldsymbol{h}^{(1)} \mid \tilde{\boldsymbol{w}}^{(1)}, \mathcal{D}) \times \ln \sum_{\boldsymbol{h}^{(2)}} P^{(2)}_{\mathrm{RBM}}(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)} \mid \boldsymbol{w}^{(2)})$$

with respect to $\boldsymbol{w}^{(2)}$. Let the solution for training the second RBM be represented by $\tilde{\boldsymbol{w}}^{(2)}$.

Then, we train $\boldsymbol{w}^{(3)}$ using the maximum likelihood estimate of the third RBM, $P^{(3)}_{\mathrm{RBM}}(\boldsymbol{h}^{(2)}, \boldsymbol{h}^{(3)} \mid \boldsymbol{w}^{(3)})$ (RBM3 in Fig.2(c)). During training, we use the feature points on the second hidden layer, which are generated by

$$P_{\mathrm{fp}}(\boldsymbol{h}^{(2)} \mid \tilde{\boldsymbol{w}}^{(1)}, \tilde{\boldsymbol{w}}^{(2)}, \mathcal{D}) := \sum_{\boldsymbol{v}, \boldsymbol{h}^{(1)}} P^{(2)}_{\mathrm{RBM}}(\boldsymbol{h}^{(2)} \mid \boldsymbol{h}^{(1)}, \tilde{\boldsymbol{w}}^{(2)}) P^{(1)}_{\mathrm{RBM}}(\boldsymbol{h}^{(1)} \mid \boldsymbol{v}, \tilde{\boldsymbol{w}}^{(1)}) \times Q_{\mathcal{D}}(\boldsymbol{v}),$$

as the pseudo-data points. In the same manner, $\boldsymbol{w}^{(r)}$ in an upper layer is trained by maximizing the log-likelihood function

$$L^{(r)}_{\mathrm{RBM}}(\boldsymbol{w}^{(r)}; \tilde{\boldsymbol{\theta}}_{r-1}) := \sum_{\boldsymbol{h}^{(r-1)}} P_{\mathrm{fp}}(\boldsymbol{h}^{(r-1)} \mid \tilde{\boldsymbol{\theta}}_{r-1}, \mathcal{D}) \times \ln \sum_{\boldsymbol{h}^{(r)}} P^{(r)}_{\mathrm{RBM}}(\boldsymbol{h}^{(r-1)}, \boldsymbol{h}^{(r)} \mid \boldsymbol{w}^{(r)}) \tag{6}$$

with respect to $\boldsymbol{w}^{(r)}$, where $\tilde{\boldsymbol{\theta}}_r := \{\tilde{\boldsymbol{w}}^{(k)} \mid k = 1, 2, \ldots, r\}$ and

$$P_{\mathrm{fp}}(\boldsymbol{h}^{(r)} \mid \tilde{\boldsymbol{\theta}}_r, \mathcal{D}) := \sum_{\boldsymbol{v}} \sum_{\boldsymbol{H}_{r-1}} Q_{\mathcal{D}}(\boldsymbol{v}) \times \prod_{k=0}^{r-1} P^{(k+1)}_{\mathrm{RBM}}(\boldsymbol{h}^{(k+1)} \mid \boldsymbol{h}^{(k)}, \tilde{\boldsymbol{w}}^{(k+1)}), \tag{7}$$

for $1 \leq r \leq R-1$. Here, $\boldsymbol{H}_r := \{\boldsymbol{h}^{(k)} \mid k = 1, 2, \ldots, r\}$, and $\tilde{\boldsymbol{w}}^{(k)}$ is the solution for training the kth RBM. Note that we define $\boldsymbol{H}_0 = \emptyset$ here. Therefore, the summation over $\boldsymbol{H}_{r-1}$ disappears when $r = 1$ in Eq. (7).

4 Interpretation of Pretraining as a Variational Approximation of MLE

The log-likelihood function of the DBM in Eq. (3) can be decomposed as

$$L_{\mathrm{DBM}}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{v}} Q_{\mathcal{D}}(\boldsymbol{v}) F_{\mathcal{H}|\mathcal{V}}(\boldsymbol{\theta}, \boldsymbol{v}) + F_{\mathrm{DBM}}(\boldsymbol{\theta}), \tag{8}$$

where $F_{\mathrm{DBM}}(\boldsymbol{\theta}) := -\ln Z_{\mathrm{DBM}}(\boldsymbol{\theta})$ is the free energy of the DBM, referred to as *the full-free energy* (FFE), and

$$F_{\mathcal{H}|\mathcal{V}}(\boldsymbol{\theta}, \boldsymbol{v}) := -\ln \sum_{\boldsymbol{H}} \exp\big(-E_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H}; \boldsymbol{\theta})\big) \tag{9}$$

is the free energy of the DBM for a specific $\boldsymbol{v}$, referred to as *the clamped-free energy* (CFE). In the following, the two types of free energy, the FFE and CFE, are approximated by different types of variational approximations, respectively.

4.1 Variational Approximation for the Clamped-Free Energy using a Deep Bayesian Network

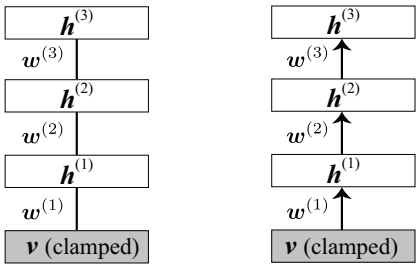

Figure 3: (a) True conditional distribution $\pi^*(\boldsymbol{H} \mid \boldsymbol{v}) = P_{\mathrm{DBM}}(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta})$. (b) Bayesian network $\pi^\dagger(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta})$ as the approximation of (a).

For the test distribution $\pi(\boldsymbol{H} \mid \boldsymbol{v})$, let us define the variational free energy as

$$\mathcal{F}_{\mathcal{H}|\mathcal{V}}[\pi] := \sum_{\boldsymbol{H}} E_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H}; \boldsymbol{\theta}) \pi(\boldsymbol{H} \mid \boldsymbol{v}) - \mathcal{H}[\pi], \quad (10)$$

where $\mathcal{H}[p] := -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \ln p(\boldsymbol{x})$ is the entropy functional. The minimum of the variational free energy under the normalization constraint for $\pi(\boldsymbol{H} \mid \boldsymbol{v})$ coincides with the CFE in Eq. (9):

$$F_{\mathcal{H}|\mathcal{V}}(\boldsymbol{\theta}, \boldsymbol{v}) = \min_{\pi} \Big\{ \mathcal{F}_{\mathcal{H}|\mathcal{V}}[\pi] \mid \sum_{\boldsymbol{H}} \pi(\boldsymbol{H} \mid \boldsymbol{v}) = 1 \Big\},$$

and $\pi(\boldsymbol{H} \mid \boldsymbol{v})$, which minimize the variational free energy, is

$$\pi^*(\boldsymbol{H} \mid \boldsymbol{v}) = P_{\mathrm{DBM}}(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta}), \quad (11)$$

where $P_{\mathrm{DBM}}(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta})$ is the true conditional distribution of the DBM conditioned with the visible layer (Fig. 3(a)).

Here, let us prepare the Bayesian network (BN) (see Fig. 3(b)) defined by

$$\pi^\dagger(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta}) := \prod_{r=0}^{R-1} P_{\mathrm{RBM}}^{(r+1)}(\boldsymbol{h}^{(r+1)} \mid \boldsymbol{h}^{(r)}, \boldsymbol{w}^{(r+1)}) \quad (12)$$

as the approximation of the true distribution in Eq. (11). The conditional distributions in the BN are the conditional distributions of the RBMs in Eq. (4). By substituting Eq. (12) into the variational free energy in Eq. (10), we obtain

$$\begin{aligned} \mathcal{F}_{\mathcal{H}|\mathcal{V}}[\pi^\dagger] &= -\ln \sum_{\boldsymbol{h}^{(1)}} \exp\big(- E_{\mathrm{RBM}}^{(1)}(\boldsymbol{v}, \boldsymbol{h}^{(1)}; \boldsymbol{w}^{(1)}) \big) \\ &- \sum_{r=1}^{R-1} \sum_{\boldsymbol{h}^{(r)}} P_{\mathrm{post}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{\theta}_r, \boldsymbol{v}) \\ &\times \ln \sum_{\boldsymbol{h}^{(r+1)}} \exp\big(- E_{\mathrm{RBM}}^{(r+1)}(\boldsymbol{h}^{(r)}, \boldsymbol{h}^{(r+1)}; \boldsymbol{w}^{(r+1)}) \big), \end{aligned} \quad (13)$$

where $\boldsymbol{\theta}_r := \{\boldsymbol{w}^{(k)} \mid k = 1, 2, \ldots, r\}$, and

$$\begin{aligned} &P_{\mathrm{post}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{\theta}_r, \boldsymbol{v}) \\ &:= \sum_{\boldsymbol{H}_{r-1}} \prod_{k=0}^{r-1} P_{\mathrm{RBM}}^{(k+1)}(\boldsymbol{h}^{(k+1)} \mid \boldsymbol{h}^{(k)}, \boldsymbol{w}^{(k+1)}), \end{aligned} \quad (14)$$

for $1 \le r \le R-1$, is the marginal distribution of the BN. Because the BN in Eq. (12) satisfies the normalization constraint $\sum_{\boldsymbol{H}} \pi^\dagger(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta}) = 1$, the BN can be a candidate of the solution to the variational minimization of Eq. (10). However, in general, $\pi^\dagger(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta}) \neq P_{\mathrm{DBM}}(\boldsymbol{H} \mid \boldsymbol{v}, \boldsymbol{\theta})$; therefore, we find the inequality

$$\mathcal{F}_{\mathcal{H}|\mathcal{V}}[\pi^\dagger] \ge \mathcal{F}_{\mathcal{H}|\mathcal{V}}[\pi^*] = F_{\mathcal{H}|\mathcal{V}}(\boldsymbol{\theta}, \boldsymbol{v}) \quad (15)$$

holds, because $\mathcal{F}_{\mathcal{H}|\mathcal{V}}[\pi^*]$ is the true minimum of the variational free energy in Eq. (10).

4.2 Variational Approximation for the Full-Free Energy using the Bethe Free Energy

If each layer in the DBM is regarded as a block, the structure of the DBM can be read as the one-dimensional Markov random field of the blocks, e.g., see Fig. 4(a). Therefore, by using its marginal distributions, we can express the DBM in Eq. (2) as

$$\begin{aligned} &P_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H} \mid \boldsymbol{\theta}) \\ &= P_{\mathrm{DBM}}(\boldsymbol{v} \mid \boldsymbol{\theta}) \prod_{r=0}^{R-1} P_{\mathrm{DBM}}(\boldsymbol{h}^{(r+1)} \mid \boldsymbol{h}^{(r)} \boldsymbol{\theta}) \\ &= P_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{h}^{(1)} \mid \boldsymbol{\theta}) \prod_{r=1}^{R-1} \frac{P_{\mathrm{DBM}}(\boldsymbol{h}^{(r)}, \boldsymbol{h}^{(r+1)} \mid \boldsymbol{\theta})}{P_{\mathrm{DBM}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{\theta})}. \end{aligned} \quad (16)$$

Following Eq. (16), let us define the variational free energy as

$$\mathcal{F}_{\mathrm{DBM}}[\phi] := \sum_{\boldsymbol{v}, \boldsymbol{H}} E_{\mathrm{DBM}}(\boldsymbol{v}, \boldsymbol{H}; \boldsymbol{\theta}) \phi(\boldsymbol{v}, \boldsymbol{H}) - \mathcal{H}[\phi], \quad (17)$$

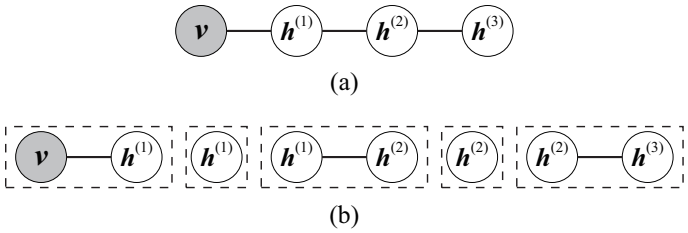

Figure 4: (a) DBM as a one-dimensional chain. (b) Bethe-type decomposition based on the CVM.

together with the test distribution $\phi(\boldsymbol{v},\boldsymbol{h})$, which is expressed by

$$\phi(\boldsymbol{v},\boldsymbol{H}) := \Gamma_{0,1}(\boldsymbol{v},\boldsymbol{h}^{(1)}) \prod_{r=1}^{R-1} \frac{\Gamma_{r,r+1}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)})}{\gamma_r(\boldsymbol{h}^{(r)})}, \tag{18}$$

where $\Gamma_{r,r+1}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)})$ is the joint distribution of $\boldsymbol{h}^{(r)}$ and $\boldsymbol{h}^{(r+1)}$, and $\gamma_r(\boldsymbol{h}^{(r)})$ is the distribution of $\boldsymbol{h}^{(r)}$. By using Eq. (18), we can express Eq. (17) as Eq. (19). Eq. (19) is identified with the (blocked) Bethe free energy of $F_{\mathrm{DBM}}(\boldsymbol{\theta})$ obtained by the cluster variation method (CVM) (Yedidia et al., 2005; Pelizzola, 2005) by decomposing the clusters according to Fig. 4(b). The variational minimization of the Bethe free energy with respect to $\boldsymbol{\gamma} = \{\gamma_r(\boldsymbol{h}^{(r)}) \mid 1 \leq r \leq R-1\}$ and $\boldsymbol{\Gamma} = \{\Gamma_{r,r+1}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)}) \mid 0 \leq r \leq R-1\}$ under the normalizing constraints,

$$\sum_{\boldsymbol{h}^{(r)}} \gamma_r(\boldsymbol{h}^{(r)}) = \sum_{\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)}} \Gamma_{r,r+1}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)}) = 1, \tag{20}$$

and the marginal constraints,

$$\gamma_r(\boldsymbol{h}^{(r)}) = \sum_{\boldsymbol{h}^{(r+1)}} \Gamma_{r,r+1}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)}), \tag{21}$$

$$\gamma_r(\boldsymbol{h}^{(r)}) = \sum_{\boldsymbol{h}^{(r-1)}} \Gamma_{r-1,r}(\boldsymbol{h}^{(r-1)},\boldsymbol{h}^{(r)}), \tag{22}$$

coincides with the FFE

$$F_{\mathrm{DBM}}(\boldsymbol{\theta}) = \min_{\{\boldsymbol{\gamma},\boldsymbol{\Gamma}\}} \big\{\mathcal{F}_{\mathrm{DBM}}[\phi] \mid \text{constraints in (20)–(22)}\big\}. \tag{23}$$

This equality originates from the exactness of the Bethe approximation (Bethe, 1935), which is the same as belief propagation (Pearl, 1988), in one-dimensional chain systems (Welling and Teh, 2003; Pelizzola, 2005).

Here, as an approximation, we neglect the marginal constraints in Eqs. (21) and (22) in the variational minimization in Eq. (23), and then, we define a new quantity as

$$F^{\dagger}_{\mathrm{DBM}}(\boldsymbol{\theta}) := \min_{\{\boldsymbol{\gamma},\boldsymbol{\Gamma}\}} \big\{\mathcal{F}_{\mathrm{DBM}}[\phi] \mid \text{constraints in (20)}\big\}. \tag{24}$$

Obviously, the inequality

$$F^{\dagger}_{\mathrm{DBM}}(\boldsymbol{\theta}) \leq F_{\mathrm{DBM}}(\boldsymbol{\theta}) \tag{25}$$

holds because the number of constraints in the variational minimization in Eq. (24) is less than that in Eq. (23). By performing the variational minimization in Eq. (24), we obtain

$$F^{\dagger}_{\mathrm{DBM}}(\boldsymbol{\theta}) = -\sum_{r=1}^{R} \ln Z^{(r)}_{\mathrm{RBM}}(\boldsymbol{w}^{(r)}), \tag{26}$$

where $Z^{(r)}_{\mathrm{RBM}}(\boldsymbol{w}^{(r)})$ is the partition function of the RBM in Eq. (4). In the derivation of Eq. (26), the equality

$$\min_{\gamma_r} \Big\{\mathcal{H}[\gamma_r] \mid \sum_{\boldsymbol{h}^{(r)}} \gamma_r(\boldsymbol{h}^{(r)}) = 1\Big\} = 0$$

is used.

4.3 Derivation of the DBN Type of Pretraining for DBMs

By using Eqs. (15) and (25), a lower bound of the log-likelihood function of the DBM in Eq. (8) is expressed as

$$L_{\mathrm{DBM}}(\boldsymbol{\theta}) \geq L^{\mathrm{lower}}_{\mathrm{DBM}}(\boldsymbol{\theta}) := -\sum_{\boldsymbol{v}} Q_{\mathcal{D}}(\boldsymbol{v}) \mathcal{F}_{H|V}[\pi^{\dagger}] + F^{\dagger}_{\mathrm{DBM}}(\boldsymbol{\theta}). \tag{27}$$

From Eqs. (5), (6), (13), and (26), we obtain

$$L^{\mathrm{lower}}_{\mathrm{DBM}}(\boldsymbol{\theta}) = L^{(1)}_{\mathrm{RBM}}(\boldsymbol{w}^{(1)}) + \sum_{r=2}^{R} L^{(r)}_{\mathrm{RBM}}(\boldsymbol{w}^{(r)};\boldsymbol{\theta}_{r-1}), \tag{28}$$

where the equality

$$P_{\mathrm{fp}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{\theta}_r, \mathcal{D}) = \sum_{\boldsymbol{v}} Q_{\mathcal{D}}(\boldsymbol{v}) P_{\mathrm{post}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{\theta}_r, \boldsymbol{v})$$

is used (cf. Eqs. (7) and (14)).

$$\mathcal{F}_{\mathrm{DBM}}[\phi] = \sum_{r=0}^{R-1} \sum_{\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)}} E_{\mathrm{RBM}}^{(r+1)}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)};\boldsymbol{w}^{(r+1)})\Gamma_{r,r+1}(\boldsymbol{h}^{(r)},\boldsymbol{h}^{(r+1)}) - \sum_{r=0}^{R-1}\mathcal{H}[\Gamma_{r,r+1}] + \sum_{r=1}^{R-1}\mathcal{H}[\gamma_r]. \quad (19)$$

Maximization of the lower bound in Eq. (28) with respect to all the parameters remains difficult, because $P_{\mathrm{fp}}(\boldsymbol{h}^{(r)} \mid \boldsymbol{\theta}_r, \mathcal{D})$ depends on all the parameters in the lower layers. Hence, we take a greedy maximization strategy in which we successively maximize the lower bound starting from the bottom layer, namely,

$$\operatorname*{gmax}_{\boldsymbol{\theta}} L_{\mathrm{DBM}}^{\mathrm{lower}}(\boldsymbol{\theta}) := \max_{\boldsymbol{w}^{(1)}} L_{\mathrm{RBM}}^{(1)}(\boldsymbol{w}^{(1)}) + \sum_{r=2}^{R} \max_{\boldsymbol{w}^{(r)}} L_{\mathrm{RBM}}^{(r)}(\boldsymbol{w}^{(r)};\tilde{\boldsymbol{\theta}}_{r-1}), \quad (29)$$

where $\tilde{\boldsymbol{\theta}}_r = \{\tilde{\boldsymbol{w}}^{(k)} \mid k = 1, 2, \ldots, r\}$ and

$$\tilde{\boldsymbol{w}}^{(r)} = \begin{cases} \operatorname{argmax}_{\boldsymbol{w}^{(1)}} L_{\mathrm{RBM}}^{(1)}(\boldsymbol{w}^{(1)}) & (r=1) \\ \operatorname{argmax}_{\boldsymbol{w}^{(r)}} L_{\mathrm{RBM}}^{(r)}(\boldsymbol{w}^{(r)};\tilde{\boldsymbol{\theta}}_{r-1}) & (r \geq 2) \end{cases}.$$

The procedure of the greedy maximization in Eq.(29) is the same as the pretraining procedure described in Sec. 3. Obviously, the inequality

$$\max_{\boldsymbol{\theta}} L_{\mathrm{DBM}}^{\mathrm{lower}}(\boldsymbol{\theta}) \geq \operatorname*{gmax}_{\boldsymbol{\theta}} L_{\mathrm{DBM}}^{\mathrm{lower}}(\boldsymbol{\theta}) \quad (30)$$

holds. From this inequality and Eq. (27), we have

$$\max_{\boldsymbol{\theta}} L_{\mathrm{DBM}}(\boldsymbol{\theta}) \geq \max_{\boldsymbol{\theta}} L_{\mathrm{DBM}}^{\mathrm{lower}}(\boldsymbol{\theta}) \geq \operatorname*{gmax}_{\boldsymbol{\theta}} L_{\mathrm{DBM}}^{\mathrm{lower}}(\boldsymbol{\theta}). \quad (31)$$

From Eq. (31), we find that the pretraining described in Sec. 3 greedily improves the lower bound of the true log-likelihood function.

The arguments in this section showed that two facts are ensured in general DBMs with discrete variables. One fact is that the variational approximation of the MLE procedure can lead to the procedure of the DBN type of pretraining for a DBM, and the second is that the pretraining procedure improves the variational bound of the true log-likelihood function of the DBM. The first fact seems to be a particularly important finding for pretraining, because it probably constitutes the first attempt to understand pretraining as a certain approximation of MLE.

In the approximation for the FFE in Sec. 4.2, we neglect the two facts: (a) the consistency between two different RBMs, that is, the marginal distributions over the upper layer in the lower RBM and over the lower layer in the upper RBM are the same (cf. Eqs. (21) and (22)) and, as a result of the approximation in (a), (b) the consistency for the double-counted entropies (the last term of Eq. (19)). By the approximation, all the RBMs are completely decoupled (cf. Eq. (26)). Here, let us consider the pretraining procedure presented in Eq. (29) in an identical case in which all the RBMs can perfectly reconstruct the distribution of the observed data and the distribution of the feature points. After the pretraining, since all the RBMs perfectly reconstruct the distribution of the observed data and that of the feature points, consistency (a), which is neglected in the approximation, is recovered. It can be understood as follows. In this case, the marginal distribution $\sum_{\boldsymbol{h}^{(1)}} \Gamma_{0,1}(\boldsymbol{v},\boldsymbol{h}^{(1)})$ is equivalent to the empirical distribution $Q_{\mathcal{D}}(\boldsymbol{v})$, and therefore, the distribution $P_{\mathrm{fp}}(\boldsymbol{h}^{(1)} \mid \tilde{\boldsymbol{w}}^{(1)}, \mathcal{D})$ is equivalent to the marginal distribution $\sum_{\boldsymbol{v}} \Gamma_{0,1}(\boldsymbol{v},\boldsymbol{h}^{(1)})$. Since the second RBM perfectly reconstructs the distribution $P_{\mathrm{fp}}(\boldsymbol{h}^{(1)} \mid \tilde{\boldsymbol{w}}^{(1)}, \mathcal{D})$, the marginal distribution over the lower layer in the second RBM, $\sum_{\boldsymbol{h}^{(2)}} \Gamma_{1,2}(\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)})$, is equivalent to $P_{\mathrm{fp}}(\boldsymbol{h}^{(1)} \mid \tilde{\boldsymbol{w}}^{(1)}, \mathcal{D})$. This means that the marginal distribution over the upper layer in the first RBM and that over the lower layer in the second RBM are equivalent. This argument can be recursively applied to the upper RBMs. The above discussion shows that the neglected consistencies (the consistency (a)) are partially recovered during the pretraining procedure.

Under a specific assumption for the structure, it is guaranteed that additive stacked RBMs, which deepen the architecture of the model, improve the variational bound when pretraining for DBNs (Hinton et al., 2006). However, the results in this section do not involve such a statement for DBMs.

5 Original Pretraining Procedure for DBMs

In the original pretraining procedure for DBMs proposed by Salakhutdinov and Hinton (2009), first, we extend the original DBM in Eq. (2) as

$$\begin{aligned} &P_{\mathrm{ex}}(\boldsymbol{v}_1,\boldsymbol{v}_2,\boldsymbol{H}_{R-1},\boldsymbol{h}_1^{(R)},\boldsymbol{h}_2^{(R)} \mid \boldsymbol{\theta}) \\ &\propto P_{\mathrm{DBM}}(\boldsymbol{v}_1,\boldsymbol{H}_{R-1},\boldsymbol{h}_1^{(R)} \mid \boldsymbol{\theta}) \\ &\quad \times P_{\mathrm{DBM}}(\boldsymbol{v}_2,\boldsymbol{H}_{R-1},\boldsymbol{h}_2^{(R)} \mid \boldsymbol{\theta}), \end{aligned} \quad (32)$$

and after the extension, we perform the DBN type of pretraining for the extended model. An example of the extension is shown in Fig. 5. The log-likelihood

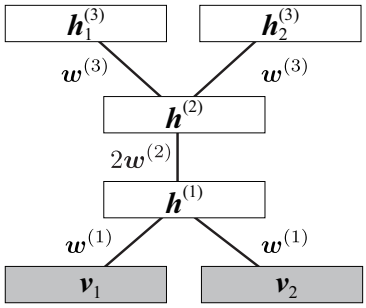

Figure 5: Extension model of the DBM in Fig. 1.

function of the extended model is

$$L_{\rm ex}(\boldsymbol{\theta}) := \sum_{\boldsymbol{v}_1,\boldsymbol{v}_2} Q_{\mathcal{D}}^{\rm ex}(\boldsymbol{v}_1,\boldsymbol{v}_2)\ln P_{\rm ex}(\boldsymbol{v}_1,\boldsymbol{v}_2 \mid \boldsymbol{\theta}), \quad (33)$$

where $P_{\rm ex}(\boldsymbol{v}_1,\boldsymbol{v}_2 \mid \boldsymbol{\theta})$ is the marginal distribution of the extended model and $Q_{\mathcal{D}}^{\rm ex}(\boldsymbol{v}_1,\boldsymbol{v}_2) := Q_{\mathcal{D}}(\boldsymbol{v}_1)\delta(\boldsymbol{v}_1,\boldsymbol{v}_2)$ is the extended empirical distribution.

Because the extended model in Eq. (32) is also the DBM, we can apply the theoretical results obtained in the previous section. From Eq. (28), the variational lower bound of the log-likelihood function in Eq. (33) is derived as

$$L_{\rm ex}^{\rm lower}(\boldsymbol{\theta}) := L_{\rm ex}^{(1)}(\boldsymbol{w}^{(1)}) + \sum_{r=2}^{R-1} L_{\rm RBM}^{(r)}(2\boldsymbol{w}^{(r)}; 2\boldsymbol{\theta}_{r-1}) + L_{\rm ex}^{(R)}(\boldsymbol{w}^{(R)}; 2\boldsymbol{\theta}_{R-1}), \quad (34)$$

where

$$L_{\rm ex}^{(1)}(\boldsymbol{w}^{(1)}) := \sum_{\boldsymbol{v}_1,\boldsymbol{v}_2} Q_{\mathcal{D}}^{\rm ex}(\boldsymbol{v}_1,\boldsymbol{v}_2)\ln \mathcal{P}_{\rm ex}^{(1)}(\boldsymbol{v}_1,\boldsymbol{v}_2 \mid \boldsymbol{w}^{(1)}),$$

$$L_{\rm ex}^{(R)}(\boldsymbol{w}^{(R)}; 2\boldsymbol{\theta}_{R-1}) := \sum_{\boldsymbol{h}^{(R-1)}} P_{\rm fp}(\boldsymbol{h}^{(R-1)} \mid 2\boldsymbol{\theta}_{R-1}, \mathcal{D}) \times \ln \mathcal{P}_{\rm ex}^{(R)}(\boldsymbol{h}^{(R-1)} \mid \boldsymbol{w}^{(R)}).$$

Here, $\mathcal{P}_{\rm ex}^{(1)}(\boldsymbol{v}_1,\boldsymbol{v}_2 \mid \boldsymbol{w}^{(1)})$ and $\mathcal{P}_{\rm ex}^{(R)}(\boldsymbol{h}^{(R-1)} \mid \boldsymbol{w}^{(R)})$ are the distributions defined by

$$\mathcal{P}_{\rm ex}^{(1)}(\boldsymbol{v}_1,\boldsymbol{v}_2 \mid \boldsymbol{w}^{(1)}) \propto \sum_{\boldsymbol{h}^{(1)}} P_{\rm RBM}^{(1)}(\boldsymbol{v}_1,\boldsymbol{h}^{(1)} \mid \boldsymbol{w}^{(1)}) P_{\rm RBM}^{(1)}(\boldsymbol{v}_2,\boldsymbol{h}^{(1)} \mid \boldsymbol{w}^{(1)}),$$

$$\mathcal{P}_{\rm ex}^{(R)}(\boldsymbol{h}^{(R-1)} \mid \boldsymbol{w}^{(R)}) \propto \Big[\sum_{\boldsymbol{h}^{(R)}} P_{\rm RBM}^{(R)}(\boldsymbol{h}^{(R-1)},\boldsymbol{h}^{(R)} \mid \boldsymbol{w}^{(R)})\Big]^2.$$

From Eq. (27), the inequality

$$L_{\rm ex}(\boldsymbol{\theta}) \geq L_{\rm ex}^{\rm lower}(\boldsymbol{\theta}) \quad (35)$$

is ensured. The original pretraining procedure for the DBM corresponds to greedily improving Eq. (34), which is the variational lower bound of the log-likelihood function in Eq. (33).

A theoretical relationship between the variational bound in Eq. (34) and the true log-likelihood function, which is the most important issue that we wish to resolve, has not been revealed. In the following, we numerically compare the log-likelihoods with their variational lower bounds. Let us consider a DBM with R hidden layers in which each layer randomly takes two or three variables and each $w_{ij}^{(r)}$ is independently drawn from the Gaussian distribution $\mathcal{N}(0,\sigma^2)$. All the variables in the DBM are the binary variable: $v_i, h_j^{(r)} \in \{0,1\}$.

In the DBM, we numerically evaluated the true log-likelihood, the variational lower bound of the true log-likelihood, the log-likelihood of the extended DBM, and the variational lower bound of the log-likelihood of the extended DBM. In the evaluation, $N = 1000$ data points, the elements of which randomly take 0 or 1, were used. For a fair comparison, the log-likelihoods were divided by the number of visible variables, namely, they are the log-likelihoods per data. Fig. 6 shows the log-likelihoods for various R. In the figure, "true," "ext," "lower bound (true)," and "lower bound (ext)" show $L_{\rm DBM}(\boldsymbol{\theta})/|\mathcal{V}|$, $L_{\rm ex}(\boldsymbol{\theta})/(2|\mathcal{V}|)$, $L_{\rm DBM}^{\rm lower}(\boldsymbol{\theta})/|\mathcal{V}|$, and $L_{\rm ex}^{\rm lower}(\boldsymbol{\theta})/(2|\mathcal{V}|)$, respectively. In the figure, $L_{\rm ex}^{\rm lower}(\boldsymbol{\theta})$ is always closer to the true log-likelihood than $L_{\rm DBM}^{\rm lower}(\boldsymbol{\theta})$. Thus, from the numerical results, we can expect $L_{\rm ex}^{\rm lower}(\boldsymbol{\theta})$ to be a better lower bound of the true log-likelihood than $L_{\rm DBM}^{\rm lower}(\boldsymbol{\theta})$:

$$L_{\rm DBM}(\boldsymbol{\theta}) \geq \frac{L_{\rm ex}^{\rm lower}(\boldsymbol{\theta})}{2} \geq L_{\rm DBM}^{\rm lower}(\boldsymbol{\theta}). \quad (36)$$

This could imply that the original pretraining procedure is more effective than the DBN type of pretraining described in Sec. 3. Note that Eq.(36) is not the theoretical result but the prediction obtained from the numerical results.

It was noted that adding layers to a DBM yields diminishing improvements in a certain variational bound (Salakhutdinov and Hinton, 2012). A similar finding can be obtained in the presented results. From Eq. (28), it is ensured that an additive layer decreases the lower bound $L_{\rm DBM}^{\rm lower}(\boldsymbol{\theta})$, because a log-likelihood function for discrete variables is non-positive. In fact, in Fig. 6, the difference between $L_{\rm DBM}(\boldsymbol{\theta})$ and $L_{\rm DBM}^{\rm lower}(\boldsymbol{\theta})$ rapidly increases with the increase in the number of hidden layers. This would suggest that an additive layer impairs the performance of the pretraining in DBMs.

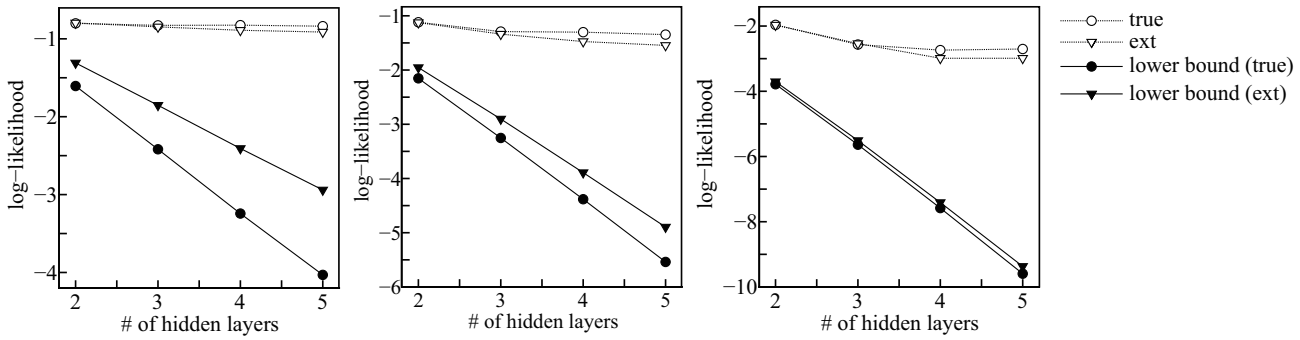

Figure 6: Log-likelihoods for various R. The left, center, and right panels show the log-likelihoods when $\sigma = 1$, 2, and 4, respectively. Each plot is the average over 1000 trials.

6 Conclusion

In the first part of this paper, the DBN type of pretraining for DBMs was discussed. The true log-likelihood function of a DBM can be decomposed into two different free energies, the CFE and FFE. By approximating the two free energies by using the different types of variational approximation, the replacing approximation by using a BN for the CFE and the Bethe-type of approximation based on the CVM for the FFE, respectively, the pretraining procedure is derived. We saw that the pretraining procedure greedily improves the variational lower bound of the true log-likelihood function. The theoretical results can be applied to general DBMs with discrete variables and will help deepen our understanding of deep learning.

In the latter part, we applied the obtained theoretical results to the original pretraining procedure for a DBM and demonstrated that the original procedure greedily improves the variational lower bound of the log-likelihood function for the extended DBM. Moreover, we numerically compared the log-likelihoods with their variational lower bounds. In the numerical results, we observed that the variational bound that the original pretraining procedure optimizes is a tighter bound of the true log-likelihood function than one that the DBN type of pretraining optimizes, and that the variational bounds rapidly depart from the true log-likelihood function with the increase in the number of hidden layers.

A theoretical connection between the true log-likelihood function and the variational bound that the original pretraining procedure optimizes, such as Eq. (36), is not revealed. It should be considered in future studies. More effective pretraining procedures for DBMs were proposed by several researchers (Hinton and Salakhutdinov, 2012; Cho et al., 2013). The Analysis of these procedures on the basis of the results obtained in this paper is also an important future issue.

Acknowledgements

This work was partially supported by CREST, Japan Science and Technology Agency and by JSPS KAKENHI Grant Numbers 15K00330, 25280089, and 15H03699.

References

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief net. *Neural Computation*, 18(7):1527–1554, 2006.

Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. *In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, pages 448–455, 2009.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 8(14):1771–1800, 2002.

T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. *In Proceedings of the 25th International Conference on Machine Learning (ICML2008)*, pages 1064–1071, 2008.

R. Salakhutdinov and H. Larochelle. Efficient learning of deep Boltzmann machines. *In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, pages 693–700, 2010.

R. Salakhutdinov and G. E. Hinton. An efficient learning procedure for deep Boltzmann machines. *Neural Computation*, 24(8):1967–2006, 2012.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.

A. Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General*, 38(33): R309–R339, 2005.

H. A. Bethe. Statistical theory of superlattices. *In Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 150:552–575, 1935.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd ed.)*. San Francisco, CA: Morgan Kaufmann, 1988.

M. Welling and Y. W. Teh. Approximate inference in Boltzmann machines. *Artificial Intelligence*, 143(1): 19–50, 2003.

G. E. Hinton and R. Salakhutdinov. A better way to pretrain deep Boltzmann machines. *In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS2012)*, pages 2447–2455, 2012.

K. Cho, T. Raiko, A. Ilin, and J. Karhunen. A two-stage pretraining algorithm for deep Boltzmann machines. *In Proceedings of the 23rd International Conference on Artificial Neural Networks (ICANN2013)*, pages 106–113, 2013.