Behaviormetrika
https://doi.org/10.1007/s41237-024-00230-9

ORIGINAL PAPER

Gaussian-discrete restricted Boltzmann machine with sparse-regularized hidden layer

Muneki Yasuda[1] · **Kaiji Sekimoto**[1]

Received: 25 December 2023 / Accepted: 26 March 2024

Abstract
Overfitting is a critical concern in machine learning, particularly when the representation capabilities of learning models surpass the complexities present in the training datasets. To mitigate overfitting, curtailing the representation power of the model through suitable techniques such as regularization is necessary. In this study, a sparse-regularization method for Gaussian–Discrete restricted Boltzmann machines (GDRBMs) is considered. A GDRBM is a variant of restricted Boltzmann machines that comprises a continuous visible layer and discrete hidden layer. In the proposed model, sparse GDRBM (S-GDRBM), a sparse prior that encourages sparse representations of the hidden layer is employed. The strength of the prior (i.e., the sparse-regularization strength) can be tuned within the standard scenario of maximum likelihood learning; that is, the strength can be adaptively tuned based on the complexities of the datasets during training. We validated the proposed S-GDRBM using numerical experiments.

Keywords Restricted Boltzmann machine · Sparse regularization · Spatial Monte Carlo integration method

1 Introduction

Overfitting is a critical issue in machine learning, and it becomes severe as the representation power of the learning model increases and the size of training dataset decreases. Regularization techniques are the most popular methods for mitigating the problem of overfitting (Bishop 2006). In standard regularization methods, such as L_1 or L_2 regularizations, penalties are imposed on the learning parameters by

Communicated by Joe Suzuki.

✉ Muneki Yasuda
muneki1005@gmail.com

[1] Graduate School of Science and Engineering, Yamagata University, Jonan 4-3-16, Yonezawa, Yamagata 992-8510, Japan

Published online: 16 April 2024

adding penalty terms to the objective functions such as loss or log likelihood functions. These standard regularization methods often involve hyperparameters (e.g., regularization coefficients) that control the strength of penalties, and the values of the hyperparameters are fixed during training.

Discriminative restricted Boltzmann machine (dRBM) is a probabilistic three-layered neural network, consisting of input, hidden, and output layers, designed for solving classification problems (Larochelle and Bengio 2008; Larochelle et al. 2012). The dRBM is constructed based on restricted Boltzmann machine (RBM) (Smolensky 1986; Hinton 2002). The representational capacity of the dRBM can be regulated by adjusting the size of the hidden layer, which expands as the hidden layer's size increases. A regularization method for the dRBM, sparse dRBM (S-dRBM), was previously proposed (Yasuda and Katsumata 2023). In this regularization method, a sparse prior for the hidden layer is employed in the form of a Laplace-type distribution, the effect of which encourages sparse representations of the hidden layer. An advantage of this regularization method is that the regularization strength (i.e., the strength of the prior) is trainable; in other words, the regularization strength can be adaptively tuned based on dataset complexity, within the standard scenario of maximum likelihood (ML) learning.

Gaussian–Bernoulli RBM (GBRBM) is a variant of RBM that can handle continuous data points (Hinton and Salakhutdinov 2006; Cho et al. 2011), and canonicalized GBRBM is a reparameterized version of the GBRBM (Yasuda and Xiong 2023). RBMs are also actively investigated in the field of physics (Decelle and Furtlehner 2021; Chen et al. 2018; Nomura and Imada 2021; Torlai et al. 2018; Carleo and Troyer 2017.) In standard GBRBMs, the hidden variables take binary values, for example, $\{0, 1\}$ or $\{-1, 1\}$. In this study, we consider (canonicalized) Gaussian–Discrete RBM (GDRBM) in which the hidden variables can accept multiple discrete values. When the hidden variables are binary, the GDRBM is equivalent to the canonicalized GBRBM. This study proposes a sparse-regularized GDRBM, referred to as sparse GDRBM (S-GDRBM), by applying the successful regularization method employed in S-dRBM.

The remainder of this paper is organized as follows. The GDRBM is defined in section 2. Section 3 presents the S-GDRBM. The S-GDRBM is obtained by combining a Laplace-type sparse prior for the hidden layer with the GDRBM. The details of the S-GDRBM are discussed in section 3.1, and the maximum-likelihood learning of the S-GDRBM based on spatial Monte Carlo integration (SMCI) method (Yasuda 2015; Yasuda and Uchizawa 2021) is discussed in section 3.2. In section 4, we demonstrate learning experiments using artificial datasets, which show that the proposed S-GDRBM effectively suppresses overfitting. Section 5 concludes the paper and presents future research directions.

2 Gaussian-discrete restricted Boltzmann machine

We consider a GDRBM defined on a complete bipartite graph consisting of two layers: visible and hidden layers. The visible layer consists of continuous visible variables $\boldsymbol{v} := \{v_i \in \mathbb{R} \mid i \in V\}$, and the hidden layer consists of discrete hidden

variables $\boldsymbol{h} := \{h_j \in \mathcal{X}_H \mid j \in H\}$, where V and H are the sets of indices of the visible and hidden variables, respectively; $\mathcal{X}_H$ is a discrete sample space. The sizes of the visible and hidden layers are denoted by n and m, respectively (i.e., $|V| = n$ and $|H| = m$). The energy function of the GDRBM is defined by

$$E_\theta(\boldsymbol{v}, \boldsymbol{h}) := \sum_{i \in V} \frac{v_i^2}{2 \operatorname{sfp} \sigma_i} - \sum_{i \in V} b_i v_i - \sum_{j \in H} c_j h_j - \sum_{i \in V} \sum_{j \in H} w_{i,j} v_i h_j, \tag{1}$$

where $\operatorname{sfp} z := \ln(1 + e^z)$ is the softplus function; here, the learning parameters, $\{b_i, \sigma_i, c_j, w_{i,j}\}$, are collectively denoted by θ. The GDRBM is a joint distribution expressed as

$$P_\theta(\boldsymbol{v}, \boldsymbol{h}) := \frac{1}{Z_\theta} \exp\big(-E_\theta(\boldsymbol{v}, \boldsymbol{h})\big), \tag{2}$$

where

$$Z_\theta := \int_{-\infty}^{+\infty} \Big(\sum_{\boldsymbol{h}} \exp\big(-E_\theta(\boldsymbol{v}, \boldsymbol{h})\big) \Big) d\boldsymbol{v}$$

is the normalization constant (or the partition function); here, $\sum_{\boldsymbol{h}}$ denotes the multiple summation over $\boldsymbol{h} \in \mathcal{X}_H^m$, and $\int_{-\infty}^{+\infty} d\boldsymbol{v}$ denotes the multiple integration over $\boldsymbol{v} \in \mathbb{R}^n$. The GDRBM in equation (2) is a generalized model of the canonicalized GBRBM (Yasuda and Xiong 2023). When the hidden variables are binary, $\mathcal{X}_H = \{0, 1\}$, the GDRBM is equivalent to the canonicalized GBRBM. The softplus function in the first term of equation (1) is employed for learning stability (Yasuda and Xiong 2023).

3 Proposed model: sparse GDRBM

The representation power of the GDRBM increases with an increase in m (i.e., the size of the hidden layer), and the overfitting problem increases in severity as the representation power increases. Therefore, the optimization of m is critical to prevent overfitting. However, in the standard scenario, m is a hyperparameter and is not trainable. Numerous studies have addressed this issue, introducing various approaches such as sparse RBM (S-RBM) (Lee et al. 2007), sparse group RBM (SG-RBM) (Luo et al. 2011), Gaussian cardinality RBM (GC-RBM) (Wan et al. 2015), and energy-function-constraint sparse RBM (ES-RBM) (Wei et al. 2019). The S-RBM, SG-RBM, GC-RBM, and ES-RBM introduce regularizers to encourage sparse representations of the hidden layer; however, they have hyperparameters related to the strength of the regularizers.

In the dRBM, an alternative sparse regularization, S-dRBM, is proposed (Yasuda and Katsumata 2023); this model introduces a regularizer that penalizes the activations of hidden variables in its energy function, aiming to encourage sparse representations of the hidden layer. The concept of the S-dRBM is similar to that of the ES-RBM. However, it has no hyperparameters, which means that the strength of regularization in

the S-dRBM can be adaptively tuned to the complexity of the dataset within the standard scenario of the ML learning (and the S-dRBM is confirmed to be superior to the ES-RBM) (Yasuda and Katsumata 2023). This section presents the S-GDRBM, which is based on the S-dRBM.

3.1 Model definition

The key concept of the proposed sparse regularization is simple; if h_j always takes zero value (i.e., h_j is always in the off-state), the influence of the variable is effectively eliminated from the model. Based on this, a sparsity assumption, similar to that in L_1 regularization, is imposed on the values of the hidden variables. In the Bayesian interpretation, L_1 regularization can be viewed as a Laplace prior (Bishop 2006; Rish and Grabarnik 2014). Here, we assume that $\mathcal{X}_H$ is a discrete sample space defined by

$$\mathcal{X}_H = \mathcal{X}_H(R) := \{-1 + r/R \mid r = 0, 1, 2, \ldots, 2R\}, \tag{3}$$

where R is a finite positive integer greater than zero; therefore, e.g., $\mathcal{X}_H(1) = \{-1, 0, 1\}$ and $\mathcal{X}_H(2) = \{-1, -1/2, 0, 1/2, 1\}$. We consider a (discrete-type) Laplace distribution over $\boldsymbol{h}$:

$$P_{\text{lap}}(\boldsymbol{h} \mid \boldsymbol{\alpha}) \propto \prod_{j \in H} \exp\Big(-(\operatorname{sfp} \alpha_j)|h_j|\Big). \tag{4}$$

In this distribution, the hidden variables are zero with high probabilities, and $\boldsymbol{\alpha} := \{\alpha_j \in \mathbb{R} \mid j \in H\}$ controls the probabilities. By combining the Laplace distribution with the GDRBM, a new model can be defined as $P_\phi(\boldsymbol{v}, \boldsymbol{h}) \propto P_\theta(\boldsymbol{v}, \boldsymbol{h}) P_{\text{lap}}(\boldsymbol{h} \mid \boldsymbol{\alpha})$; thus, the resultant model is expressed as

$$P_\phi(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z_\phi} \exp\Big(-E_\theta(\boldsymbol{v}, \boldsymbol{h}) - \sum_{j \in H} (\operatorname{sfp} \alpha_j)|h_j|\Big), \tag{5}$$

where

$$Z_\phi := \int_{-\infty}^{+\infty} \left\{ \sum_{\boldsymbol{h}} \exp\Big(-E_\theta(\boldsymbol{v}, \boldsymbol{h}) - \sum_{j \in H} (\operatorname{sfp} \alpha_j)|h_j|\Big) \right\} d\boldsymbol{v}$$

is the normalization constant, and ϕ denotes the set of parameters comprising θ and $\boldsymbol{\alpha}$. The second term in the exponent of equation (5) functions as the penalties for non-zero hidden variables, and the strength of the penalties is controlled by $\boldsymbol{\alpha}$. Equation (5) is the S-GDRBM. The S-GDRBM is identical to the GDRBM when $\alpha_j \to -\infty$ (i.e., $\operatorname{sfp} \alpha_j = 0$) for all $j \in H$.

The layer-wise conditional distributions of the S-GDRBM are as follows:

$$P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i \in V} \frac{1}{\sqrt{2\pi \operatorname{sfp} \sigma_i}} \exp\left\{ -\frac{(v_i - \lambda_i(\boldsymbol{h}))^2}{2 \operatorname{sfp} \sigma_i} \right\}, \tag{6}$$

$$P_\phi(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_{j \in H} \frac{\exp(\tau_j(\boldsymbol{v}) h_j - (\operatorname{sfp} \alpha_j)|h_j|)}{G_j(\boldsymbol{v})}, \tag{7}$$

where

$$\tau_j(\boldsymbol{v}) := c_j + \sum_{i \in V} w_{i,j} v_i, \quad \lambda_i(\boldsymbol{h}) := (\operatorname{sfp} \sigma_i)\left(b_i + \sum_{j \in H} w_{i,j} h_j \right), \tag{8}$$

and

$$G_j(\boldsymbol{v}) := \sum_{h_j} \exp\left(\tau_j(\boldsymbol{v}) h_j - (\operatorname{sfp} \alpha_j)|h_j|\right) \tag{9}$$

is the normalization constant of $P_\phi(h_j \mid \boldsymbol{v})$. The marginal distribution over $\boldsymbol{v}$ is obtained as

$$P_\phi(\boldsymbol{v}) = \frac{1}{Z_\phi} \exp\left(- \sum_{i \in V} \frac{v_i^2}{2 \operatorname{sfp} \sigma_i} + \sum_{i \in V} b_i v_i + \sum_{j \in H} \ln G_j(\boldsymbol{v}) \right). \tag{10}$$

The marginal distribution over $\boldsymbol{h}$ is obtained through the multivariate Gaussian integral, which leads to

$$P_\phi(\boldsymbol{h}) = \frac{1}{\mathcal{Z}_\phi} \exp\left(\boldsymbol{\beta}^{\mathrm{t}} \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^{\mathrm{t}} \boldsymbol{J} \boldsymbol{h} - \sum_{j \in H} (\operatorname{sfp} \alpha_j)|h_j| \right), \tag{11}$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ and $\boldsymbol{J} \in \mathbb{R}^{m \times m}$ are defined as

$$\boldsymbol{\beta} := \boldsymbol{c} + \boldsymbol{W}^{\mathrm{t}} \boldsymbol{S} \boldsymbol{b}, \quad \boldsymbol{J} := \boldsymbol{W}^{\mathrm{t}} \boldsymbol{S} \boldsymbol{W}, \tag{12}$$

where $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose (i, i)-element is $\operatorname{sfp} \sigma_i$, $\boldsymbol{b} \in \mathbb{R}^n$ and $\boldsymbol{c} \in \mathbb{R}^m$ are the vectors of b_i and c_j, respectively, and $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ is the matrix of $w_{i,j}$; $\mathcal{Z}_\phi$ is the normalization constant defined by

$$\mathcal{Z}_\phi := \sum_{\boldsymbol{h}} \exp\left(\boldsymbol{\beta}^{\mathrm{t}} \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^{\mathrm{t}} \boldsymbol{J} \boldsymbol{h} - \sum_{j \in H} (\operatorname{sfp} \alpha_j)|h_j| \right), \tag{13}$$

which is expressed in terms of Z_ϕ as

$$\mathcal{Z}_\phi = Z_\phi \exp\left(-\frac{1}{2} \sum_{i \in V} \ln(2\pi \operatorname{sfp} \sigma_i) - \frac{1}{2} \boldsymbol{b}^{\mathrm{t}} \boldsymbol{S} \boldsymbol{b} \right). \tag{14}$$

Equation (11) can be regarded as a Boltzmann machine defined on a fully connected graph:

$$P_\phi(\boldsymbol{h}) \propto \exp\left(\sum_{j\in H} q_j(h_j) + \sum_{i<j\in H} J_{i,j}h_i h_j\right),$$

where $q_j(h_j) := \beta_j h_j + J_{j,j}h_j^2/2 - (\operatorname{sfp}\alpha_j)|h_j|$ are the potential on the hidden variables.

The marginal distribution of the S-GDRBM (as well as the GDRBM) over $\boldsymbol{v}$ can be viewed as a Gaussian mixture model. The marginal distribution is expressed as

$$P_\phi(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) P_\phi(\boldsymbol{h}).$$

Here, the conditional distribution, $P_\phi(\boldsymbol{v} \mid \boldsymbol{h})$, is the Gaussian distribution (cf. equation (6)); thus, this expression can be considered a Gaussian mixture model with $|\mathcal{X}_H|^m$ Gaussian components in which $P_\phi(\boldsymbol{h})$ functions as the mixture weight. The number of Gaussian components rises exponentially with increasing m and power-functionally with increasing R because $|\mathcal{X}_H|^m = (2R+1)^m$. Therefore, although it is small compared to the increase of m, the increase of R may also increase the representation power of the S-GDRBM.

3.2 Maximum-likelihood learning based on spatial Monte Carlo integration

We assume that a training dataset consisting of N data points, $D := \{\mathbf{v}^{(\mu)}\}_{\mu=1}^N$, is obtained. The learning of the S-GDRBM is achieved by maximizing the log likelihood,

$$\ell(\phi) := \frac{1}{N}\sum_{\mu=1}^N \ln P_\phi(\mathbf{v}^{(\mu)}), \tag{15}$$

with respect to ϕ. From equation (10), the log likelihood is expressed as

$$\ell(\phi) = -\sum_{i\in V}\frac{1}{2\operatorname{sfp}\sigma_i}\mathbb{E}_D[v_i^2] + \sum_{i\in V} b_i\mathbb{E}_D[v_i] + \sum_{j\in H}\mathbb{E}_D\left[\ln G_j(\boldsymbol{v})\right] - \ln Z_\phi,$$

where $\mathbb{E}_D[\cdots]$ denotes the sample average over the training dataset, that is,

$$\mathbb{E}_D[f(\boldsymbol{v})] = \frac{1}{N}\sum_{\mu=1}^N f\left(\mathbf{v}^{(\mu)}\right).$$

Therefore, the gradients of the log likelihood are obtained as follows. The gradients for b_i and σ_i are

$$\frac{\partial\ell(\phi)}{\partial b_i} = \mathbb{E}_D[v_i] - \mathbb{E}_\phi[v_i] \tag{16}$$

and

$$\frac{\partial \ell(\phi)}{\partial \sigma_i} = \frac{\operatorname{sig} \sigma_i}{2(\operatorname{sfp} \sigma_i)^2} \big(\mathbb{E}_D[v_i^2] - \mathbb{E}_\phi[v_i^2] \big), \tag{17}$$

respectively, where $\operatorname{sig} z := 1/(1+e^{-z})$ is the sigmoid function, and $\mathbb{E}_\phi[\cdots]$ denotes the model expectation of the S-GDRBM, that is,

$$\mathbb{E}_\phi[\cdots] := \int_{-\infty}^{+\infty} \sum_{\boldsymbol{h}} (\cdots) P_\phi(\boldsymbol{v}, \boldsymbol{h}) d\boldsymbol{v}.$$

Next, the gradients for c_j and $w_{i,j}$ are

$$\frac{\partial \ell(\phi)}{\partial c_j} = \mathbb{E}_D\big[H_j(\boldsymbol{v})\big] - \mathbb{E}_\phi[h_j] \tag{18}$$

and

$$\frac{\partial \ell(\phi)}{\partial w_{i,j}} = \mathbb{E}_D\big[v_i H_j(\boldsymbol{v})\big] - \mathbb{E}_\phi[v_i h_j], \tag{19}$$

respectively, where

$$H_j(\boldsymbol{v}) := \sum_{h_j} h_j P_\phi(h_j \mid \boldsymbol{v}) = \frac{\sum_{h_j} h_j \exp(\tau_j(\boldsymbol{v}) h_j - (\operatorname{sfp} \alpha_j)|h_j|)}{G_j(\boldsymbol{v})}. \tag{20}$$

Finally, the gradients for α_j is

$$\frac{\partial \ell(\phi)}{\partial \alpha_j} = (\operatorname{sig} \alpha_j)\big(-\mathbb{E}_D\big[Q_j(\boldsymbol{v})\big] + \mathbb{E}_\phi\big[|h_j|\big] \big), \tag{21}$$

where

$$Q_j(\boldsymbol{v}) := \sum_{h_j} |h_j| P_\phi(h_j \mid \boldsymbol{v}) = \frac{\sum_{h_j} |h_j| \exp(\tau_j(\boldsymbol{v}) h_j - (\operatorname{sfp} \alpha_j)|h_j|)}{G_j(\boldsymbol{v})}. \tag{22}$$

The ML learning is conducted using a gradient ascent method based on the gradients in equations (16), (17), (18), (19), and (21), which implies that the sparsity parameters $\boldsymbol{\alpha}$ and the other learning parameters θ are simultaneously tuned within the ML learning. To encourage sparsity, relatively large values are preferred for the initial values of $\boldsymbol{\alpha}$, for example $\alpha_j \approx 10$, as recommended in reference (Yasuda and Katsumata 2023). However, these gradients include the intractable model expectations, the computational costs of which exponentially grow with the size of the model (the model expectations can be computed when m is sufficiently small; see Appendix A for the details).

In the following, an approximation of the model expectations based on the first-order SMCI method (Yasuda 2015; Yasuda and Uchizawa 2021) (which can be viewed as a Rao-Backwellization) is considered; SMCI-based evaluation has outperformed the evaluation based on the standard Monte Carlo integration (MCI)

in Bernoulli–Bernoulli RBMs (Sekimoto and Yasuda 2023) and deep Boltzmann machines (Katsumata and Yasuda 2021). We assume that we have K sample points, $S := \{\mathbf{v}^{(\nu)}, \mathbf{h}^{(\nu)}\}_{\nu=1}^{K}$, drawn from the S-GDRBM. Here, the first-order SMCI method is briefly introduced. The visible and hidden variables are collectively denoted by $\boldsymbol{x} = \boldsymbol{v} \cup \boldsymbol{h}$, and νth sample point is denoted by $\mathbf{x}^{(\nu)} = \mathbf{v}^{(\nu)} \cup \mathbf{h}^{(\nu)}$. Based on the first-order SMCI method, the model expectation for a function of $\boldsymbol{x}_t \subseteq \boldsymbol{x}$ is evaluated as

$$\mathbb{E}_{\phi}[f(\boldsymbol{x}_t)] \approx \mathbb{E}_S\left[\sum_{\boldsymbol{x}_t} f(\boldsymbol{x}_t) P_{\phi}(\boldsymbol{x}_t \mid \boldsymbol{x}_{\partial t})\right] = \frac{1}{K}\sum_{\nu=1}^{K}\sum_{\boldsymbol{x}_t} f(\boldsymbol{x}_t) P_{\phi}(\boldsymbol{x}_t \mid \mathbf{x}_{\partial t}^{(\nu)}), \tag{23}$$

where $\boldsymbol{x}_{\partial t} \subseteq \boldsymbol{x}$ is the nearest-neighbor variables of $\boldsymbol{x}_t$, for example, $\boldsymbol{x}_{\partial t} = \boldsymbol{h}$ when $\boldsymbol{x}_t = \{v_i\}$ and $\boldsymbol{x}_{\partial t} = \boldsymbol{x} \setminus \{v_i, h_j\}$ when $\boldsymbol{x}_t = \{v_i, h_j\}$. Here, $\mathbb{E}_S[\cdots]$ denotes the sample average over the sample set S. In equation (23), the sum over $x_i \in \boldsymbol{x}_t$ is replaced with the integration over x_i when x_i is continuous. Based on equation (23), the model expectations, $\mathbb{E}_{\phi}[v_i]$ and $\mathbb{E}_{\phi}[v_i^2]$, are approximated as

$$\mathbb{E}_{\phi}[v_i] \approx \frac{1}{K}\sum_{\nu=1}^{K}\int_{-\infty}^{+\infty} v_i P_{\phi}(v_i \mid \mathbf{h}^{(\nu)}) = \frac{1}{K}\sum_{\nu=1}^{K}\lambda_i(\mathbf{h}^{(\nu)}) \tag{24}$$

and

$$\mathbb{E}_{\phi}[v_i^2] \approx \frac{1}{K}\sum_{\nu=1}^{K}\int_{-\infty}^{+\infty} v_i^2 P_{\phi}(v_i \mid \mathbf{h}^{(\nu)}) = \operatorname{sfp}\sigma_i + \frac{1}{K}\sum_{\nu=1}^{K}\lambda_i(\mathbf{h}^{(\nu)})^2, \tag{25}$$

respectively, where equation (6) is used. Similarly, using equation (7), the model expectations, $\mathbb{E}_{\phi}[h_j]$ and $\mathbb{E}_{\phi}[|h_j|]$, are approximated as

$$\mathbb{E}_{\phi}[h_j] \approx \frac{1}{K}\sum_{\nu=1}^{K}\sum_{h_j} h_j P_{\phi}(h_j \mid \mathbf{v}^{(\nu)}) = \frac{1}{K}\sum_{\nu=1}^{K} H_j(\mathbf{v}^{(\nu)}) \tag{26}$$

and

$$\mathbb{E}_{\phi}\left[|h_j|\right] \approx \frac{1}{K}\sum_{\nu=1}^{K}\sum_{h_j} |h_j| P_{\phi}(h_j \mid \mathbf{v}^{(\nu)}) = \frac{1}{K}\sum_{\nu=1}^{K} Q_j(\mathbf{v}^{(\nu)}), \tag{27}$$

respectively. Finally, the approximation of $\mathbb{E}_{\phi}[v_i h_j]$ is considered. Based on equation (23), it is approximated as

$$\mathbb{E}_{\phi}[v_i h_j] \approx \frac{1}{K}\sum_{\nu=1}^{K}\int_{-\infty}^{+\infty}\sum_{h_j} v_i h_j P_{\phi}(v_i, h_j \mid \mathbf{v}_{-i}^{(\nu)}, \mathbf{h}_{-j}^{(\nu)}) dv_i, \tag{28}$$

where $\boldsymbol{v}_{-i} := \boldsymbol{v} \setminus \{v_i\}$ and $\boldsymbol{h}_{-j} := \boldsymbol{h} \setminus \{h_j\}$. The conditional distribution in the right hand side of equation (28) is

$$P_\phi(v_i, h_j \mid \boldsymbol{v}_{-i}, \boldsymbol{h}_{-j}) \propto \exp\left(-\frac{v_i^2}{2\,\mathrm{sfp}\,\sigma_i} + b_{i,j}(\boldsymbol{h}_{-j})v_i + c_{j,i}(\boldsymbol{v}_{-i})h_j - (\mathrm{sfp}\,\alpha_j)|h_j| + w_{i,j}v_i h_j\right), \tag{29}$$

where

$$b_{i,j}(\boldsymbol{h}_{-j}) := b_i + \sum_{\ell \in H\setminus\{j\}} w_{i,\ell} h_\ell = \frac{\lambda_i(\boldsymbol{h})}{\mathrm{sfp}\,\sigma_i} - w_{i,j}h_j,$$
$$c_{j,i}(\boldsymbol{v}_{-i}) := c_j + \sum_{k \in V\setminus\{i\}} w_{k,j} v_k = \tau_j(\boldsymbol{v}) - w_{i,j}v_i.$$

From equations (28) and (29), we obtain

$$\mathbb{E}_\phi[v_i h_j] \approx \frac{\mathrm{sfp}\,\sigma_i}{K} \sum_{v=1}^{K} \frac{\sum_{h_j}(w_{i,j}h_j + b_{i,j}(\mathbf{h}_{-j}^{(v)}))h_j \exp(-e_j^{(v)}(h_j))}{\sum_{h_j}\exp(-e_j^{(v)}(h_j))}, \tag{30}$$

where

$$e_j^{(v)}(h_j) := -\frac{\mathrm{sfp}\,\sigma_i}{2} w_{i,j}^2 h_j^2 - \left(c_{j,i}(\mathbf{v}_{-i}^{(v)}) + (\mathrm{sfp}\,\sigma_i) b_{i,j}(\mathbf{h}_{-j}^{(v)}) w_{i,j}\right) h_j + (\mathrm{sfp}\,\alpha_j)|h_j|.$$

By substituting the model expectations, $\mathbb{E}_\phi[\cdots]$, in the gradients in equations (16), (17), (18), (19), and (21) with the corresponding approximations provided in equations (24), (25), (26), (27), and (30), respectively, the approximated gradients are obtained. The cost of the computation of the SMCI-based expectations is $O(Knm)$; thus, they can be computed even when the size of the S-GDRBM is large. Although the aforementioned SMCI-based approximations are formulated for the S-GDRBM, they can be directly applied to the GDRBM by setting $\mathrm{sfp}\,\alpha_j = 0$ (i.e., $\alpha_j \to -\infty$).

To demonstrate the validity of the SMCI-based evaluation, using numerical experiments, we compared the approximation accuracy of it with that of the MCI-based evaluation on small-sized S-GDRBMs with $n = m = 10$. The parameter setup of the S-GDRBMs was as follows: the bias parameters, $\boldsymbol{b}$ and $\boldsymbol{c}$, and weight parameters, $\boldsymbol{W}$, were drawn from a uniform distribution in the interval $[-\beta, \beta]$, $\boldsymbol{\alpha}$ were drawn from a uniform distribution in the interval $[-10, 10]$, and $\{\sigma_i\}$ were fixed by $\sigma_i = \ln(e-1)$ (i.e., $\mathrm{sfp}\,\sigma_i = 1$). The sample set with $K = 1000$ was generated using layer-wised blocked Gibbs sampling on the S-GDRBM. Figure 1 depicts the mean absolute errors (MAEs) between the exact model expectations and their approximations. Because the S-GDRBMs are small, the exact model expectations can be evaluated (see Appendix A). The SMCI-based evaluation outperforms the MCI-based evaluation in terms of MAE.

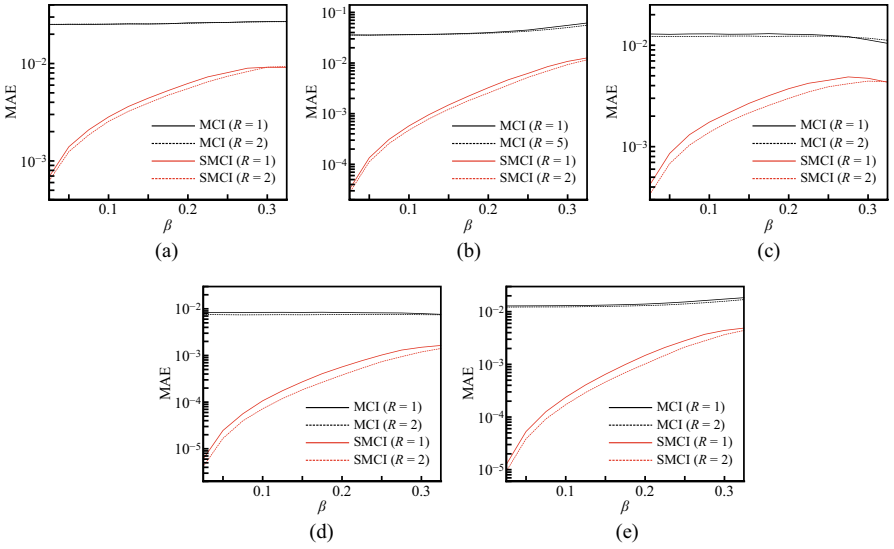

Fig. 1 MAEs between exact expectations and their approximations for various β: (a) $\mathbb{E}_\phi[v_i]$, (b) $\mathbb{E}_\phi[v_i^2]$, (c) $\mathbb{E}_\phi[h_j]$, (d) $\mathbb{E}_\phi[|h_j|]$, and (e) $\mathbb{E}_\phi[v_i h_j]$. The plots present the average values of 3000 experiments

4 Numerical experiment

In this section, we demonstrate the ML learning of the S-GDRBM and compare it to that of the GDRBM using artificial training datasets in which the artificial training datasets were generated from the GBRBM (Yasuda and Xiong 2023).

First, we demonstrate numerical experiments on small-sized models. The size of the data-generative GBRBM, $P_{\text{gen}}(\boldsymbol{v}, \boldsymbol{h})$, was $n = 5$ and $m = 3$, in which the bias parameters, $\boldsymbol{b}$ and $\boldsymbol{c}$, and weight parameters, $\boldsymbol{W}$, were drawn from a Gaussian distribution with zero mean and variance 0.01, and $\{\sigma_i\}$ were fixed by $\sigma_i = \ln(e-1)$. Using the data-generative GBRBM, artificial datasets with size N were generated based on layer-wised blocked Gibbs sampling. The use of artificial datasets is appropriate for our purpose because their complexities can be controlled, and moreover, the degree of generalization can be monitored (using a negative cross-entropy described below).

For the artificial datasets, the ML learnings were conducted using the GDRBM and S-GDRBM (with $R = 1, 2$) in which the sizes of the visible layers were $n = 5$ and the sizes of the hidden layers were $m = 3$ or $m = 7$. The bias parameters were initialized to zero, while the weight parameters were initialized using (Gaussian-type) Xavier's initialization (Glorot and Bengio 2010), and $\{\sigma_i\}$ were initialized to $\sigma_i = -3$ for all $i \in V$. In the S-GDRBM, the initial values of $\boldsymbol{\alpha}$ were set to a fixed value of $\alpha_j = 10$ for all $j \in H$. The adamax optimizer (Kingma and Ba 2015) with the full-batch training was used in the gradient ascent. The log likelihood,

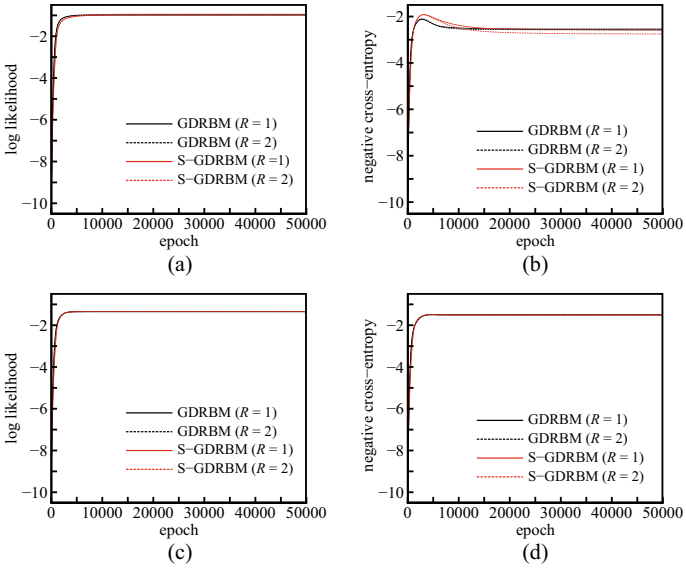

Fig. 2 Log likelihoods and negative cross-entropies obtained based on the exact learning. The sizes of the learning models are $n = 5$ and $m = 3$

$$\frac{1}{N}\sum_{\mu=1}^{N} \ln P_{\mathrm{tr}}(\mathbf{v}^{(\mu)}),$$

and the negative cross-entropy,

$$\int_{-\infty}^{+\infty} P_{\mathrm{gen}}(\boldsymbol{v}) \ln P_{\mathrm{tr}}(\boldsymbol{v}) d\boldsymbol{v},$$

were used as measures to assess the quality of the learning process, where $P_{\mathrm{tr}}(\boldsymbol{v}) = P_{\theta}(\boldsymbol{v})$ when the learning model is the GDRBM and $P_{\mathrm{tr}}(\boldsymbol{v}) = P_{\phi}(\boldsymbol{v})$ when it is the S-GDRBM. The log likelihood represents the fitness to the training dataset, and the negative cross-entropy represents the degree of generalization. As the learning proceeds without overfitting, both log likelihood and negative cross-entropy monotonically increase; whereas the negative cross-entropy decreases as overfitting begins to appear. The log likelihood and cross-entropy were exactly computed because the sizes of the data-generative and learning models were sufficiently small (see Appendix A). The exact learning and the SMCI-based learning presented in section 3.2 were conducted. In the SMCI-based learning, the sample points, S, required to evaluate the model expectations in equations (24), (25), (26), (27), and (30) were obtained based on 10-steps layer-wised blocked Gibbs sampling starting from the training data points (i.e., the sampling procedure used in CD_{10} (Hinton 2002)).

Figures 2–5 depict the values of the log likelihoods and negative cross-entropies against the training epoch; the upper plots, labeled (a) and (b), in the figures

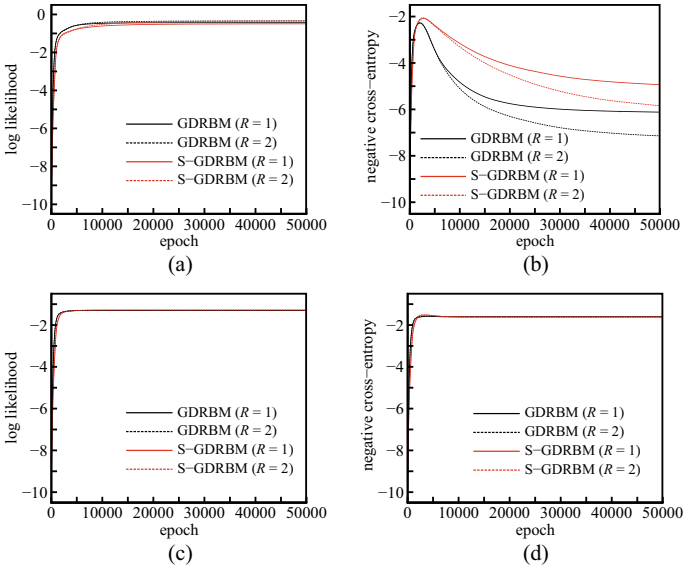

Fig. 3 Log likelihoods and negative cross-entropies obtained based on the exact learning. The sizes of the learning models are $n = 5$ and $m = 7$

represent the results obtained when $N = 10$, while the lower plots, labeled (c) and (d), display the results for the case $N = 100$. The plots in the figures present the average values obtained from 100 experiments. The results in figures 2 and 3 were obtained based on the exact learning and those in figures 4 and 5 were based on the SMCI-based learning. Overfitting is particularly observed in figures 3(b) and 5(b). We can observe that the S-GDRBMs successfully reduce overfitting. Whereas, the S-GDRBMs yield similar results to those of the GDRBMs in the experiments where overfitting is not a significant issue.

Next, we demonstrate numerical experiments on larger models in which the size of the data-generative GBRBM was $n = m = 50$. The parameter setup of the data-generative GBRBM was as follows: the bias and weight parameters were drawn from Gaussian distributions with zero mean and variances 0.05 and 0.002, respectively, and $\{\sigma_i\}$ were the same as in the aforementioned experiments. For the artificial datasets generated from the data-generative GBRBM, the SMCI-learnings (with CD_{50}) were conducted using the GDRBM and S-GDRBM (with $R = 1, 2$) in which the sizes of the visible layers were $n = 50$ and the sizes the hidden layers were $m = 50$ or $m = 100$. The initialization of the learning parameters were the same as in the aforementioned experiments, and the adamax optimizer with the mini-batch training was used in which the mini-batch size was B. Figure 6 depicts the values of negative cross-entropies against the training epoch; (a) displays the results for the learning models with $m = 50$ when $N = 1000$ and $B = 100$, and (b) displays the results for the learning models with $m = 100$ when $N = 150$ and $B = 30$. The

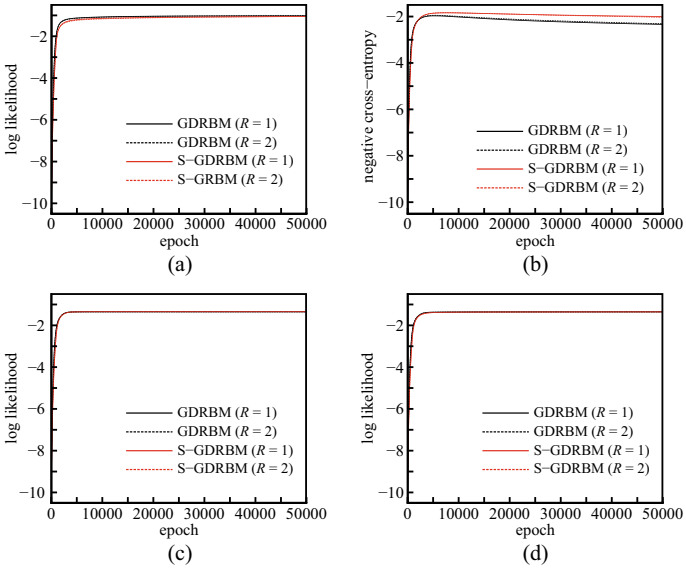

Fig. 4 Log likelihoods and negative cross-entropies obtained based on the SMCI-based learning. The sizes of the learning models are $n = 5$ and $m = 3$

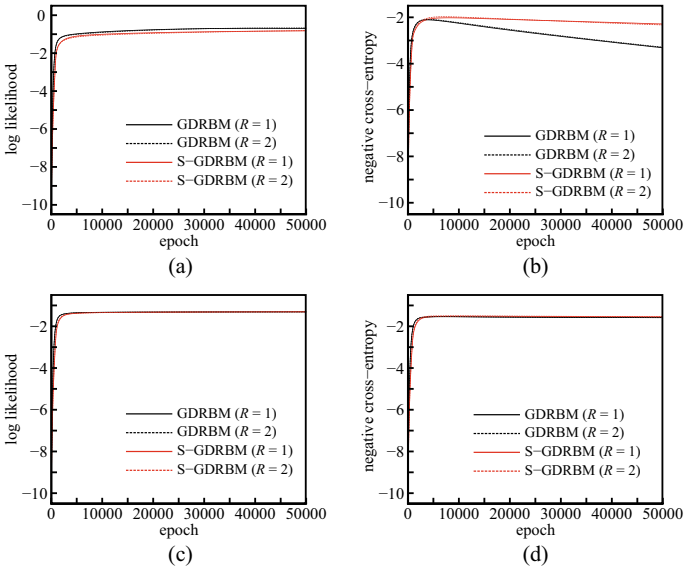

Fig. 5 Log likelihoods and negative cross-entropies obtained based on the SMCI-based learning. The sizes of the learning models are $n = 5$ and $m = 7$

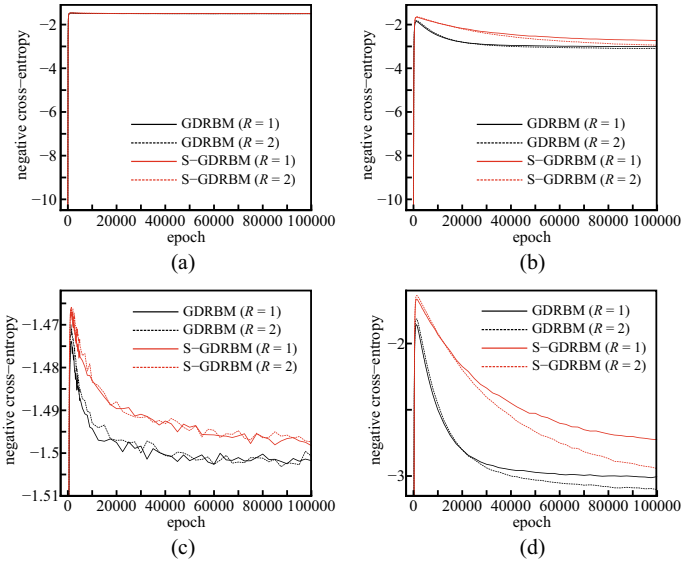

Fig. 6 Negative cross-entropies obtained based on the SMCI-based learning: (a) $m = 50$ and $N = 1000$ ($B = 100$) and (b) $m = 100$ and $N = 150$ ($B = 30$). (c) and (d) are the enlarged plots of (a) and (d), respectively. The plots in these figures present the average values of 30 experiments

negative cross-entropy was evaluated using a sampling-based approximation[1] More pronounced overfitting is observed in figure 6(b). The S-GDRBM successfully reduces overfitting; moreover, the peaks of the rise of the S-GDRBMs are higher than those of the GDRBMs (similar behaviors can be observed in figures 2(b) and 3(b)).

On the trained models obtained in the experiments in figure 6, we evaluate $\rho := \sum_{j \in H} \mathbb{E}_{\mathrm{tr}}[|h_j|]/m$, where $\mathbb{E}_{\mathrm{tr}}[\cdots]$ denotes the expectation on the trained models. $\rho \in [0, 1]$ can be read as the statistical activation-ratio of the hidden layer; $\rho = 1$ when all hidden variables always take ± 1 and $\rho = 0$ when all hidden variables always take zero. In the models with $R = 2$, the hidden variables can take two kinds of activations, $|h_j| = 1$ and $|h_j| = 1/2$, and we regard the former as the strong activation and the latter as the weak activation. If the effect of sparse regularization functions as expected, the values of ρ are suppressed. Table 1 presents the ρ-values on the trained models; here, $\mathbb{E}_{\mathrm{tr}}[|h_j|]$ was computed based on the SMCI-evaluation. The ρ-values of the S-GDRBMs are considerably lower than those of the GDRBMs, which means the proposed regularization functions. From (a) to (b) in table 1, the ρ -values of the GDRBMs increase and approach one; this implies that the effect of redundant hidden-variable-activations causes overfitting. Conversely, the ρ-values of

[1] The negative cross-entropy was evaluated based on equation (34). In the equation, $\mathbb{E}_{\phi}[v_i]$ and $\mathbb{E}_{\phi}[v_i^2]$ were evaluated using the SMCI-based evaluation in equations (24) and (25), respectively, and the normalization constant was evaluated based on marginalized annealed importance sampling (Yasuda and Takahashi 2022); the expectations in the third term in equation (34) were evaluated using equation (36) based on the standard MCI.

Table 1 Values of ρ: (a) the trained models obtained in the experiments in figure 6(a) and (b) the trained models obtained in the experiments in figure 6(b). The values of the table are the average values obtained from 30 experiments

	GDRBM		S-GDRBM	
	(a)	(b)	(a)	(b)
$R = 1$	0.875	0.990	0.235	0.101
$R = 2$	0.818	0.975	0.285	0.129

the S-GDRBMs decrease, which implies that the S-GDRBMs shrink the effect of redundant hidden-variable-activations to suppress overfitting.

We can observe that the ρ-values of the S-GDRBMs with $R = 1$ are lower than those with $R = 2$ in table 1, which implies that the effect of sparse regularization is more enhanced in the case $R = 1$. This is intuitively understood as follows: the penalties for the hidden variables taking non-zero values tend to be larger in $R = 1$ when $\boldsymbol{\alpha}$ of both S-GDRBMs are the same. In addition, compared with the S-GDRBM with $R = 1$, the S-GDRBM with $R = 2$ more significantly decreases the negative cross-entropies in figures 3(b) and 6(d). However, we consider that these results do not immediately indicate that the S-GDRBM with $R = 1$ is superior to that with $R = 2$. As mentioned in section 3.1, the representation power of the S-GDRBM can be increased by increasing the R-value. There is the possibility that the S-GDRBMs with $R = 2$ or more are more suitable for more complex training datasets than the S-GDRBM with $R = 1$. The aforementioned experimental results confirm the proposed sparse regularization functions, i.e., the strength of regularization is adaptively tuned during training. This might seem counterintuitive because the ML learning aims to achieve a good fit to the training data and does not inherently prioritize the suppression of overfitting. From the ML perspective, the strength of regularization should ideally decrease to zero (i.e., α_j goes to $-\infty$) because the solution exhibiting overfitting will be globally optimum. This matter might be considered as follows. The model learns the abstract of the data distribution in the early stage of the learning; both log likelihood and cross-entropy grow in this stage. After the early stage, the model starts to be finely tuned to learn the details of the data distribution and to increase the log likelihood. This fine tuning causes overfitting. The sparse regularization prevents the fine tuning by shrinking the hidden variables representations and attempts the model to stay at a locally optimum near the point reached in the early stage. Figure 7 depicts the long-term learning version of the experiments in figures 3(a) and 3(b). The S-GDRBMs converge to better solutions in terms of the negative cross-entropy. However, not as much as the GDRBMs, the S-GDRBMs also exhibit the tendency of overfitting. If early stopping could be properly conducted, the learning solution presenting a high negative cross-entropy can be obtained. However, appropriate early stopping in terms of the negative cross-entropy is not practical because the true data-generative model is unknown. The log likelihood for a separate test dataset may be used as the alternative criterion for early stopping. However, the log likelihood involves the intractable normalization

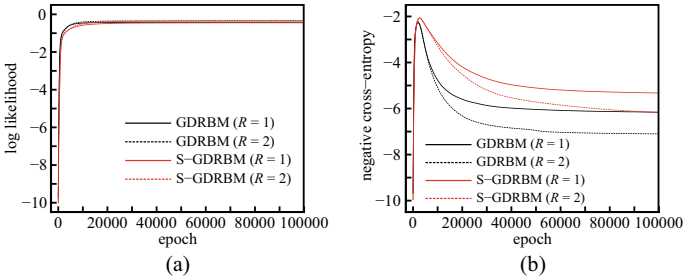

Fig. 7 Long-term learning version of the experiments in figures 3(a) and 3(b)

constant, a precise evaluation of which is expensive in large systems even if a sampling-based approximation is employed.

5 Conclusion and future studies

In this study, a sparse-regularized GDRBM, S-GDRBM, is proposed by imposing a Laplace-like prior on the hidden layer. In the S-GDRBM, the strength of sparse regularization (in other words, the strength of the prior) is trainable in contrast to that in conventional sparse regularizations. The results of our numerical experiments in section 4 show that the proposed regularization method functioned as expected. The present regularization becomes strong for training datasets in which overfitting is severe and is weakened for datasets in which overfitting is not severe, which implies the strength of regularization is adaptively tuned during training.

As mentioned in section 3, there are several related works that aim to promote sparse representations of the hidden layer (Lee et al. 2007; Luo et al. 2011; Wan et al. 2015; Wei et al. 2019). Another relevant work is infinite RBM (iRBM) (Côté and Larochelle 2016), which treats m as a random variable and tunes its distribution during training (note that the iRBM and its hybrid-type learning algorithm (Peng et al. 2018) have hyperparameters). In the iRBM, the effective size of m is optimized according to the complexity of the training dataset. The objective of the iRBM study is similar to that of the present study. The S-GDRBM and the related works (excluding the ES-RBM[2]) are not in direct competition, suggesting that the S-GDRBM can potentially be used in conjunction with them for further developments. The combination with the iRBM is important, and it will be conducted in our future studies. In the future, additional studies will explore the applications of the S-GDRBM in various contexts. These could include its use as a feature extractor (Yasuda and Xiong 2023) or as an input converter for classification systems (Kanno and Yasuda 2021).

[2] In the ES-RBM, the hidden variables being $\{0, 1\}$-binary is essential; therefore, the S-GDRBM cannot directly apply to the ES-RBM.

Evaluation of Exact Expectations on S-GDRBM

Consider a S-GDRBM and assume that m (the size of hidden layer) is sufficiently small and an expectation on $P_\phi(\boldsymbol{h})$, $\mathbb{E}_\phi^{\mathrm{h}}[\cdots] := \sum_{\boldsymbol{h}}(\cdots)P_\phi(\boldsymbol{h})$ can be evaluated by performing the multiple summation, where $P_\phi(\boldsymbol{h})$ is the marginal distribution of the S-GDRBM expressed in equation (11). In this case, the model expectations, $\mathbb{E}_\phi[v_i]$, $\mathbb{E}_\phi[v_i^2]$, $\mathbb{E}_\phi[h_j]$, $\mathbb{E}_\phi[|h_j|]$, and $\mathbb{E}_\phi[v_i h_j]$, can be obtained. The model expectations, $\mathbb{E}_\phi[h_j]$ and $\mathbb{E}_\phi[|h_j|]$, are evaluated through $\mathbb{E}_\phi[h_j] = \mathbb{E}_\phi^{\mathrm{h}}[h_j]$ and $\mathbb{E}_\phi[|h_j|] = \mathbb{E}_\phi^{\mathrm{h}}[|h_j|]$, respectively. Using $P_\phi(\boldsymbol{v}, \boldsymbol{h}) = P_\phi(\boldsymbol{v} \mid \boldsymbol{h})P_\phi(\boldsymbol{h})$, we obtain

$$\mathbb{E}_\phi[v_i] = \sum_{\boldsymbol{h}} \left(\int_{-\infty}^{+\infty} v_i P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) d\boldsymbol{v} \right) P_\phi(\boldsymbol{h}) = \sum_{\boldsymbol{h}} \lambda_i(\boldsymbol{h}) P_\phi(\boldsymbol{h}) = \mathbb{E}_\phi^{\mathrm{h}}\left[\lambda_i(\boldsymbol{h})\right], \quad (31)$$

where $\lambda_i(\boldsymbol{h})$ is defined in equation (8). In a similar manner,

$$\mathbb{E}_\phi[v_i^2] = \sum_{\boldsymbol{h}} \left(\int_{-\infty}^{+\infty} v_i^2 P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) d\boldsymbol{v} \right) P_\phi(\boldsymbol{h}) = \mathrm{sfp}\, \sigma_i + \mathbb{E}_\phi^{\mathrm{h}}\left[\lambda_i(\boldsymbol{h})^2\right], \quad (32)$$

$$\mathbb{E}_\phi[v_i h_j] = \sum_{\boldsymbol{h}} h_j \left(\int_{-\infty}^{+\infty} v_i P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) d\boldsymbol{v} \right) P_\phi(\boldsymbol{h}) = \mathbb{E}_\phi^{\mathrm{h}}\left[h_j \lambda_i(\boldsymbol{h})\right]. \quad (33)$$

are obtained.

When m is sufficiently small, the value of the log likelihood in equation (15) can be computed because the normalization constant of the S-GDRBM, Z_ϕ, can be computed using equation (14) (note that $\mathcal{Z}_\phi$ can be obtained by performing the multiple summation in equation (13)). In this situation, a negative cross-entropy between different S-GDRBMs can be evaluated. Consider a negative cross-entropy defined as

$$H_{\mathrm{cross}} := \int_{-\infty}^{+\infty} P_\phi(\boldsymbol{v}) \ln P_{\phi'}(\boldsymbol{v}) d\boldsymbol{v}.$$

The negative cross-entropy is rewritten as

$$H_{\mathrm{cross}} = -\sum_{i \in V} \frac{\mathbb{E}_\phi[v_i^2]}{2\,\mathrm{sfp}\,\sigma_i'} + \sum_{i \in V} b_i' \mathbb{E}_\phi[v_i] + \sum_{j \in H} \mathbb{E}_\phi[\ln G_j'(\boldsymbol{v})] - \ln Z_{\phi'}, \quad (34)$$

where $G_j'(\boldsymbol{v})$ denotes $G_j(\boldsymbol{v})$ with ϕ'. $\mathbb{E}_\phi[v_i]$ and $\mathbb{E}_\phi[v_i^2]$ are obtained using equations (31) and (32), respectively. The evaluation of $\mathbb{E}_\phi[\ln G_j'(\boldsymbol{v})]$ is as follows. Based on the reproductive property of Gaussian, we obtain

$$\int_{-\infty}^{+\infty} P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) \ln G_j'(\boldsymbol{v}) d\boldsymbol{v} = \int_{-\infty}^{+\infty} \mathcal{N}(z \mid \mu_j(\boldsymbol{h}), s_j^2) \ln G_j'(z) dz, \quad (35)$$

where $G_j'(z) := \ln \sum_{h_j} \exp\{(c_j' + z)h_j - (\mathrm{sfp}\,\alpha_j')|h_j|\}$; here, $\mathcal{N}(z \mid \mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2, and

$$\mu_j(\boldsymbol{h}) := \sum_{i \in V} w'_{i,j} \lambda_i(\boldsymbol{h}), \quad s_j^2 := \sum_{i \in V} (w'_{i,j})^2 \operatorname{sfp} \sigma_i.$$

Equation (35) leads to

$$\begin{aligned} \mathbb{E}_\phi[\ln G'_j(\boldsymbol{v})] &= \sum_{\boldsymbol{h}} \left(\int_{-\infty}^{+\infty} P_\phi(\boldsymbol{v} \mid \boldsymbol{h}) \ln G'_j(\boldsymbol{v}) d\boldsymbol{v} \right) P(\boldsymbol{h}) \\ &= \mathbb{E}_\phi^{\mathrm{h}} \left[\int_{-\infty}^{+\infty} \mathcal{N}(z \mid \mu_j(\boldsymbol{h}), s_j^2) \ln G'_j(z) dz \right]. \end{aligned} \tag{36}$$

Although the formulations in this appendix are obtained based on the S-GDRBM, they can be applied to GDRBMs (by setting $\operatorname{sfp} \alpha_j = 0$) and GBRBMS (by setting $\operatorname{sfp} \alpha_j = 0$ and $\mathcal{X}_H = \{0, 1\}$).

Acknowledgements We would like to thank Yuuki Yokoyama for the valuable discussions.

Funding This work was supported by JSPS KAKENHI Grant Number 21K11778 and JST, the establishment of University fellowships towards the creation of science technology innovation, Grant Number JPMJFS2104.

Data availability The datasets presented in this manuscript were generated during the study, and they are available upon reasonable request from the corresponding author. The detailed generation process of the datasets is described in the experimental section.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer-Verlag, New York

Carleo G, Troyer M (2017) Solving the quantum many-body problem with artificial neural networks. Science 355(6325):602–606

Chen J, Cheng S, Xie H, Wang L, Xiang T (2018) Equivalence of restricted boltzmann machines and tensor network states. Physical Review B 97(8):085104

Cho K, Ilin A, Raiko T (2011) Improved learning of gaussian-bernoulli restricted boltzmann machines. In Proc. of the 21th International Conference on Artificial Neural Networks pp. 10–17

Côté MA, Larochelle H (2016) An infinite restricted boltzmann machine. Neural Computation 28(7):1265–1288

Decelle A, Furtlehner C (2021) Restricted boltzmann machine: Recent advances and mean-field theory. Chinese Physics B 30(4):040202

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In Proc. of the 13th International Conference on Artificial Intelligence and Statistics 9:249–256
Hinton G (2002) Training products of experts by minimizing contrastive divergence. Neural Computation 14(8):1771–1800
Hinton G, Salakhutdinov R (2006) Reducing the dimensionality of data with neural networks. Science 313(5788):504–507
Kanno Y, Yasuda M (2021) Multi-layered discriminative restricted boltzmann machine with untrained probabilistic layer. In Proc. of the 25th International Conference on Pattern Recognition pp. 7655–7660
Katsumata T, Yasuda M (2021) Effective fine-tuning training of deep boltzmann machine based on spatial monte carlo integration. Nonlinear Theory and its Applications, IEICE 12(3):377–390
Kingma DP, Ba LJ (2015) Adam: A method for stochastic optimization. In Proc. of the 3rd International Conference on Learning Representations pp. 1–13
Larochelle H, Bengio Y (2008) Classification using discriminative restricted Boltzmann machines. In Proc. of the 25th International Conference on Machine Learning pp. 536–543
Larochelle H, Mandel M, Pascanu R, Bengio Y (2012) Learning Algorithms for the Classification Restricted Boltzmann Machine. The Journal of Machine Learning Research 13(1):643–669
Lee H, Ekanadham C, Ng AY (2007) Sparse deep belief net model for visual area V2. In Proc. of the Advances in Neural Information Processing Systems 20 pp. 873–880
Luo H, Shen R, Niu C (2011) Sparse group restricted boltzmann machines. In Proc. of the 25th AAAI Conference on Artificial Intelligence pp. 429–434
Nomura Y, Imada M (2021) Dirac-type nodal spin liquid revealed by refined quantum many-body solver using neural-network wave function, correlation ratio, and level spectroscopy. Physical Review X 11(3):031034
Peng X, Gao X, Li X (2018) On better training the infinite restricted boltzmann machines. Machine Learning 107:943–968
Rish I, Grabarnik G (2014) Sparse Modeling: Theory, Algorithms, and Applications. CRC Press
Sekimoto K, Yasuda M (2023) Effective learning algorithm for restricted boltzmann machines via spatial monte carlo integration. Nonlinear Theory and its Applications, IEICE 14(2):228–241
Smolensky P (1986) Information processing in dynamical systems: foundations of harmony theory. Parallel distributed processing: Explorations in the microstructure of cognition 1:194–281
Torlai G, Mazzola G, Carrasquilla J, Troyer M, Melko R, Carleo G (2018) Neural-network quantum state tomography. Nature Physics 14(5):447–450
Wan C, Jin X, Ding G, Shen D (2015) Gaussian cardinality restricted boltzmann machines. In Proc. of the 29th AAAI Conference on Artificial Intelligence pp. 3031–3037
Wei J, Lv J, Yi Z (2019) A new sparse restricted boltzmann machine. International Journal of Pattern Recognition and Artificial Intelligence 33(10):1951004
Yasuda M (2015) Monte carlo integration using spatial structure of markov random field. Journal of the Physical Society of Japan 84(3):034001
Yasuda M, Katsumata T (2023) Discriminative restricted boltzmann machine with trainable sparsity. Nonlinear Theory and its Applications, IEICE 14(2):207–214
Yasuda M, Takahashi C (2022) Free energy evaluation using marginalized annealed importance sampling. Physical Review E 106(2):024127
Yasuda M, Uchizawa K (2021) A generalization of spatial monte carlo integration. Neural Computation 33(4):1037–1062
Yasuda M, Xiong Z (2023) New learning algorithm of gaussian–bernoulli restricted boltzmann machine and its application in feature extraction. In Proc. of the 2023 International Symposium on Nonlinear Theory and Its Applications pp. 134–137