2023 International Symposium on Nonlinear Theory and Its Applications
NOLTA2023, September 26-29, 2023, Catania and Online

New Learning Algorithm of Gaussian–Bernoulli Restricted Boltzmann Machine and its Application in Feature Extraction

Muneki Yasuda[†] and Zhongren Xiong[‡]

†Graduate School of Science and Engineering, Yamagata University
4–3–16 Jounan, Yonezawa, Yamagata 992-8510, Japan
‡Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
Email: muneki@yz.yamagata-u.ac.jp[†]

Abstract— Feature extraction is essential in various data analyses applications. Therefore, the development of an universal feature extractor is critical. A restricted Boltzmann machine (RBM) is a powerful candidate for such an universal feature extractor. In this study, we focus on a Gaussian–Bernoulli RBM (GBRBM) and canonicalize it through re-parameterization. An effective learning algorithm for the canonicalized GBRBM is proposed based on spatial Monte Carlo integration. Using numerical experiments, we demonstrate that the GBRBM outperforms the standard denoising autoencoder as the feature extractor in strong noisy environments.

1. Introduction

Feature extraction (or more limited, data-normalization) is essential for data analyses. However, the best feature extraction method differs for each dataset. Therefore, the appropriate feature extraction method for each dataset is determined by trial and error. Hence, the development of an universal feature extractor is critical.

Restricted Boltzmann machines (RBMs) are potential candidates for such universal feature extractor. RBMs can be regarded as distribution-based autoencoders and have been successful in, e.g., dimensional reduction [1] and pre-training of deep learning [2, 3]. An RBM-based feature extractor has employed in an stochastic classification system [4], in which the feature extractor is used as the input converter, for improving noise robustness of the classification system.

In this study, we focus on a Gaussian–Bernoulli RBM (GBRBM) to treat continuous data [1, 5]. In Sec. 2, we modify the ordinary GBRBM for canonicalization; the resultant GBRBM is a canonical exponential family with respect to the learning parameters. Section 3 presents an effective learning algorithm based on spatial Monte Carlo integration (SMCI) [6, 7] for the canonicalized GBRBM. SMCI is an effective MCI-like method on Markov random fields based on a Rao-Blackwellization. SMCI-based learning has been successful in Bernoulli–Bernoulli RBMs [8]. In Sec. 4, we numerically demonstrate the validation of feature extraction based on the GBRBM trained by the proposed learning algorithm.

2. Canonicalized Gaussian–Bernoulli restricted Boltzmann machine

GBRBM, a bipartite-type of the MRF, is introduced to treat continuous data [1]. Consider that a GBRBM consists of the visible layer, having continuous visible variables $\boldsymbol{v} := \{v_i \in (-\infty, +\infty) \mid i \in V\}$, and hidden layer, having binary hidden variables $\boldsymbol{h} := \{h_j \in \{0, 1\} \mid j \in H\}$. Cho *et. al.* [5] modified the original energy function and used the following energy function:

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

where all learning parameters, $\{b_i, \sigma_i^2, c_j, w_{i,j}\}$, are collectively denoted by θ. Learning parameters $\{\sigma_i^2\}$ frequently become a cause of unstable learning; without a special treatment (e.g., a tuning of the learning rate), they may turn negative during usual gradient-based learning. To prevent this, Cho *et. al.* reparameterized σ_i^2 as $\sigma_i^2 \to \exp s_i$ [5]. However, this reparameterization can cause exponential divergence.

In this study, the energy function in equation (1) is modified. Based on Eq. (1), the resultant GBRBM is not a canonical exponential family with respect to the learning parameters (or is not a log-linear model). Reparametrizing, $b_i/\sigma_i^2 \to b_i$ and $w_{i,j}/\sigma_i^2 \to w_{i,j}$, and ignoring the constant term leads to a novel energy function:

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

This energy function let the resultant GBRBM,

$$P_\theta(\boldsymbol{v}, \boldsymbol{h}) := \frac{1}{Z_\theta} \exp\big(-E_\theta(\boldsymbol{v}, \boldsymbol{h})\big) \tag{3}$$

ORCID iDs First Author: 0000-0001-5531-9842, Second Author: 0009-0004-3999-0862 be a canonical exponential family, where Z_θ is the partition function (or the normalization constant). This canonicalization does not change the representation power. Furthermore, σ_i^2 is reparameterized as $\sigma_i^2 \to \mathrm{sfp}\, s_i$, where $\mathrm{sfp}\, x := \ln(1 + e^x)$ is the softplus function. This reparameterization does not have the risk of exponential divergence. Based on the canonicalized GBRBM in Eq. (3), the conditional distributions of each visible and hidden layers given the others are as follows:

$$P_\theta(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_{j\in H} \frac{\exp(\tau_j(\boldsymbol{v})h_j)}{1 + \exp \tau_j(\boldsymbol{v})}, \tag{4}$$

$$P_\theta(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i\in V} \frac{1}{\sqrt{2\pi\, \mathrm{sfp}\, s_i}} \exp\Big\{ -\frac{(v_i - \lambda_i(\boldsymbol{h}))^2}{2\, \mathrm{sfp}\, s_i} \Big\}, \tag{5}$$

where $\tau_j(\boldsymbol{v}) := c_j + \sum_{i\in V} w_{i,j} v_i$ and $\lambda_i(\boldsymbol{h}) := (\mathrm{sfp}\, s_i)(b_i + \sum_{j\in H} w_{i,j} h_j)$. In the rest of this paper, the term “GBRBM” denotes the canonicalized GBRBM.

The training of the GBRBM is achieved through the maximum likelihood. Given the dataset composed of N data points: $D := \{\mathbf{v}^{(\mu)} \in (-\infty, +\infty)^{|V|} \mid \mu = 1, 2, \ldots, N\}$, the log likelihood is defined as

$$\ell_\theta(D) := \frac{1}{N} \sum_{\mu=1}^{N} \ln \sum_{\boldsymbol{h}} P_\theta(\mathbf{v}^{(\mu)}, \boldsymbol{h}). \tag{6}$$

The gradients of this log likelihood are obtained as follows:

$$\frac{\partial \ell_\theta(D)}{\partial b_i} = \mathbb{E}_D[v_i] - \mathbb{E}_\theta[v_i], \tag{7}$$

$$\frac{\partial \ell_\theta(D)}{\partial s_i} = \frac{\mathrm{sig}\, s_i}{2(\mathrm{sfp}\, s_i)^2} (\mathbb{E}_D[v_i^2] - \mathbb{E}_\theta[v_i^2]), \tag{8}$$

$$\frac{\partial \ell_\theta(D)}{\partial c_j} = \mathbb{E}_D[\, \mathrm{sig}\, \tau_j(\boldsymbol{v})] - \mathbb{E}_\theta[h_j], \tag{9}$$

$$\frac{\partial \ell_\theta(D)}{\partial w_{i,j}} = \mathbb{E}_D[v_i\, \mathrm{sig}\, \tau_j(\boldsymbol{v})] - \mathbb{E}_\theta[v_i h_j], \tag{10}$$

where $\mathrm{sig}\, x := 1/(1 + e^{-x})$ is the sigmoid function, $\mathbb{E}_D[\cdots]$ denotes the sample average over the dataset D, i.e., $\mathbb{E}_D[f(\boldsymbol{v})] = N^{-1} \sum_{\mu=1}^{N} f(\mathbf{v}^{(\mu)})$, and $\mathbb{E}_\theta[\cdots] := \int_{-\infty}^{+\infty} \sum_{\boldsymbol{h}} (\cdots) P_\theta(\boldsymbol{v}, \boldsymbol{h}) d\boldsymbol{v}$ denotes the expectation of the GBRBM. The gradients in Eqs. (7)–(10) includes the intractable expectations of the GBRBM.

3. Learning Algorithm based on Spatial Monte Carlo Integration

To execute gradient-based learning, the intractable expectations, $\mathbb{E}_\theta[v_i]$, $\mathbb{E}_\theta[v_i^2]$, $\mathbb{E}_\theta[h_j]$, and $\mathbb{E}_\theta[v_i h_j]$, have to be evaluated using an approximate method. The contrastive divergence (CD) method is the most popular method [9]. In the CD method, the intractable expectations are approximated using Monte Carlo integration (MCI) (i.e., the sample average) over the sample set generated using the layer-wise blocked Gibbs sampling based on the conditional distributions in Eqs. (4) and (5), in which each Gibbs-sampling run is started from data point $\mathbf{v}^{(\mu)}$.

SMCI is an effective MCI-like method on MRFs, and it is more accurate than the standard MCI [6, 7]. In this section, the approximations of the intractable expectations based on the first-order SMCI method (i.e., SMCI in which the target and sum regions are the same) are derived. Suppose that we have the sample set composed of T sample points, $S := \{(\mathbf{v}^{(t)}, \mathbf{h}^{(t)}) \mid t = 1, 2, \ldots, T\}$, drawn from $P_\theta(\boldsymbol{v}, \boldsymbol{h})$. Based on the first-order SMCI method, $\mathbb{E}_\theta[v_i]$ is approximated as

$$\mathbb{E}_\theta[v_i] \approx \mathbb{E}_S\Big[\int_{-\infty}^{+\infty} v_i P_\theta(v_i \mid \boldsymbol{h}) dv_i \Big] = \mathbb{E}_S\left[\lambda_i(\boldsymbol{h})\right], \tag{11}$$

where Eq. (5) is used. Here, expression $\mathbb{E}_S\,[\cdots]$ denotes the sample average over the sample set S, i.e., $\mathbb{E}_S\,[f(\boldsymbol{v}, \boldsymbol{h})] = T^{-1} \sum_{t=1}^{T} f(\mathbf{v}^{(t)}, \mathbf{h}^{(t)})$. Similarly, $\mathbb{E}_\theta[v_i^2]$ is approximated as

$$\mathbb{E}_\theta[v_i^2] \approx \mathbb{E}_S\Big[\int_{-\infty}^{+\infty} v_i^2 P_\theta(v_i \mid \boldsymbol{h}) dv_i \Big] = \mathrm{sfp}\, s_i + \mathbb{E}_S\left[\lambda_i(\boldsymbol{h})^2\right]. \tag{12}$$

Using Eq. (4), $\mathbb{E}_\theta[h_j]$ is approximated as

$$\mathbb{E}_\theta[h_j] \approx \mathbb{E}_S\Big[\sum_{h_j=0,1} h_j P_\theta(h_j \mid \boldsymbol{v}) \Big] = \mathbb{E}_S\left[\, \mathrm{sig}\, \tau_j(\boldsymbol{v})\right]. \tag{13}$$

Finally, the approximation of $\mathbb{E}_\theta[v_i h_j]$ is considered. Based on the first-order SMCI method, it is approximated as

$$\mathbb{E}_\theta[v_i h_j] \approx \mathbb{E}_S\Big[\int_{-\infty}^{+\infty} \sum_{h_j=0,1} v_i h_j P_\theta(v_i, h_j \mid \boldsymbol{v}_{-i}, \boldsymbol{h}_{-j}) dv_i \Big], \tag{14}$$

where $\boldsymbol{v}_{-i} := \boldsymbol{v} \setminus \{v_i\}$ and $\boldsymbol{h}_{-j} := \boldsymbol{h} \setminus \{h_j\}$. The conditional distribution in this expression is obtained as

$$\begin{aligned} &P_\theta(v_i, h_j \mid \boldsymbol{v}_{-i}, \boldsymbol{h}_{-j}) \\ &\quad \propto \exp\Big(-\frac{v_i^2}{2\, \mathrm{sfp}\, s_i} + b_{i,j}(\boldsymbol{h}_{-j}) v_i + c_{j,i}(\boldsymbol{v}_{-i}) h_j + w_{i,j} v_i h_j \Big), \end{aligned} \tag{15}$$

where $b_{i,j}(\boldsymbol{h}_{-j}) := \lambda_i(\boldsymbol{h})/(\mathrm{sfp}\, s_i) - w_{i,j} h_j$ and $c_{j,i}(\boldsymbol{v}_{-i}) := \tau_j(\boldsymbol{v}) - w_{i,j} v_i$. Thus, from Eqs. (14) and (15), the approximation of $\mathbb{E}_\theta[v_i h_j]$ is obtained as

$$\mathbb{E}_\theta[v_i h_j] \approx (\mathrm{sfp}\, s_i) \mathbb{E}_S\Big[\big(w_{i,j} + b_{i,j}(\boldsymbol{h}_{-j})\big)\, \mathrm{sig}\, K_{i,j}(\boldsymbol{v}_{-i}, \boldsymbol{h}_{-j}) \Big], \tag{16}$$

where

$$K_{i,j}(\boldsymbol{v}_{-i}, \boldsymbol{h}_{-j}) := \frac{\mathrm{sfp}\, s_i}{2} (w_{i,j}^2 + 2 w_{i,j} b_{i,j}(\boldsymbol{h}_{-j})) + c_{j,i}(\boldsymbol{v}_{-i}).$$

The effectiveness of the proposed SMCI method, i.e., Eqs. (11), (12), (13), and (16), over the standard MCI method was confirmed via numerical experiments using small-sized GBRBMs. However, we did not display the results due to space limitation.

The proposed learning is summarized as follows. The sample set S is obtained by the same procedure as the k-step CD (CD_k) method, where k denotes the number of the blocked Gibbs sampling run before obtaining each sample point. and subsequently, using the obtained sample set the intractable expectations in the gradients in Eqs. (7)–(10), i.e., $\mathbb{E}_\theta[v_i]$, $\mathbb{E}_\theta[v_i^2]$, $\mathbb{E}_\theta[h_j]$, and $\mathbb{E}_\theta[v_i h_j]$, are evaluated based on Eqs. (11), (12), (13), and (16). The learning parameters are updated using the approximate gradients obtained the aforementioned procedure. The order of the computational time of the proposed learning is $O(N|V||H|)$ which is the same as that of the CD method.

4. Feature Extraction based on Gaussian–Bernoulli Restricted Boltzmann Machine

An RBM has an aspect as a distribution-based autoencoder. In the GBRBM, the marginal distribution over the visible layer, $P_\theta(\boldsymbol{v})$, is encoded as $P_\theta(\boldsymbol{h}) = \int_{-\infty}^{+\infty} P_\theta(\boldsymbol{h} \mid \boldsymbol{v}) P_\theta(\boldsymbol{v}) d\boldsymbol{v}$ in the hidden layer, where $P_\theta(\boldsymbol{h} \mid \boldsymbol{v})$ is the conditional distribution in Eq. (4). Therefore, the marginal distribution over the hidden layer, $P_\theta(\boldsymbol{h})$, can be considered as an encoding distribution (i.e., a feature distribution) of $P_\theta(\boldsymbol{v})$. Here, $P_\theta(\boldsymbol{v})$ is perfectly reconstructed by the decoding process: $P_\theta(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P_\theta(\boldsymbol{v} \mid \boldsymbol{h}) P_\theta(\boldsymbol{h})$, where $P_\theta(\boldsymbol{v} \mid \boldsymbol{h})$ is the conditional distribution in Eq. (5). Based on this encoding and decoding relation, we consider the feature extraction based on the GBRBM for an input $\mathbf{v}$ as

$$f_j(\mathbf{v}) := \sum_{\boldsymbol{h}} h_j P_\theta(\boldsymbol{h} \mid \mathbf{v}) = \mathrm{sig}\, \tau_j(\mathbf{v}). \tag{17}$$

The jth feature for the input is defined by the expectation of jth hidden variable.

In the following, we demonstrate the proposed feature extraction using MNIST database consisting of 60,000 training and 10,000 test data points. MNIST is a database of ten handwritten digits from "0" to "9"; each digit in the database is a 28 × 28 gray-scaled image (i.e., a 784-dimensional vector). First, using training dataset D_{train}, the GBRBM with 784 visible and 500 hidden variables was trained in an unsupervised scenario based on the SMCI training method proposed in Sec. 3. The data points, $\mathbf{v} \in \{0, 1, \ldots, 255\}^{|V|}$, in both training and test datasets were normalized by $\mathrm{v}_i \leftarrow 2(\mathrm{v}_i/255) - 1$, i.e., all elements fall in the interval $[-1, +1]$; after normalization, small noises (i.e., Gaussian noise with standard deviation $\sigma_{\text{small}} = 0.01$) were added to the training dataset. In the training, we used the stochastic gradient method with a batch size of 256 and the fixed learning rate was 0.001; the blocked Gibbs sampling same as CD_{10} was used in the sampling process.

Subsequently, we conducted feature extraction for test dataset D_{test} based on the trained GBRBM using Eq. (17): $\boldsymbol{f}(\mathbf{v})$, $\mathbf{v} \in D_{\text{test}}$. The dimension of each feature is 500. We visually verify the validation of the proposed feature extraction based on feature maps obtained by dimensionality reduction methods; a better feature is expected to be higher clustered in the feature maps. To obtain the feature maps, we used t-SNE [10] and UMAP [11]. For comparison, feature extraction based on the standard denoising autoencoder, which is the three-layered fully-connected neural network with the ReLU activation, was conducted, in which the sizes of the input and output layers were 784, and the size of the hidden layer was 500. In the training of the denoising autoencoder, Gaussian noise with standard deviation σ_{small} were used as input noise, and adam optimizer [12] with a batch size of 256 was used. In the denoising autoencoder, the feature is the output of the hidden layer for the corresponding input.

Figure 1 depicts the feature maps for the "clean" test dataset. Although all maps display well-clustered structures, the map obtained by the GBRBM appears to be the best in terms of the degree of cohesiveness of the clusters; the clusters of "5" and "8" are splitting in top panels of Figs. 1(a) and (b). Figure 2 depicts the feature maps for the "noisy" test dataset, in which the data points in the test dataset were corrupted by adding strong noise (i.e., Gaussian noise with standard deviation $\sigma = 2$). In the maps of the corrupted test dataset and of features obtained from the denoising autoencoder (i.e., Figs. 2(a) and (b)), the clustered structures were largely broken due to strong additive noise. In contrast, the clustered structure remains in the map of the features obtained from the GBRBM.

5. Conclusion

In this study, we first modified the energy function of the GBRBM to canonicalize the resultant model; the modified GBRBM belongs to the canonical exponential family with respect to the learning parameters. Next, for the modified GBRBM, an effective learning algorithm was proposed based on SMCI. Finally, the GBRBM was applied to the feature extraction problem. We numerically confirmed that the feature extractor based on the GBRBM is superior to that based on the standard denoising autoencoder in strong noisy environments.

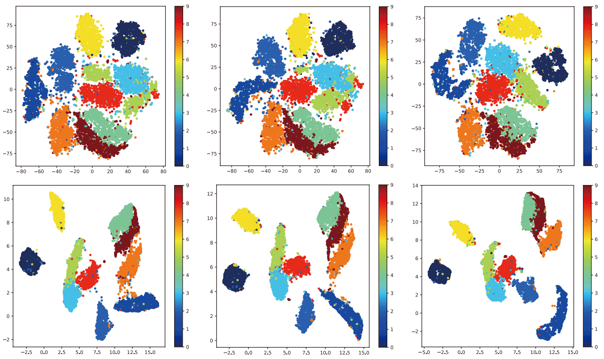
Acknowledgments

This work was partially supported by JSPS KAKENHI (Grant No. 21K11778), we would like to thank Y. Kanno and Y. Tanaka for the variable discussions.

References

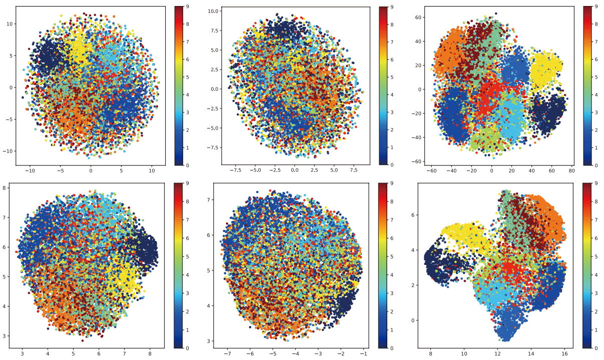
[1] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5788):504–507, 2006.

[2] G. Hinton, S. Osindero, and Y. W. Teh. A fast learning



(a) (b) (c)

Figure 1: Feature maps obtained from t-SNE (top) and UMAP (bottom): (a) test dataset, and features obtained form the test dataset based on (b) denoising autoencoder and (c) GBRBM.



(a) (b) (c)

Figure 2: Feature maps obtained from t-SNE (top) and UMAP (bottom): (a) corrupted test dataset, and features obtained form the corrupted test dataset based on (b) denoising autoencoder and (c) GBRBM.

algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[3] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. *In Proc. of the 12th International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[4] Y. Kanno and M. Yasuda. Multi-layered discriminative restricted boltzmann machine with untrained probabilistic layer. *In Proc. of the 25th International Conference on Pattern Recognition*, pages 7655–7660, 2021.

[5] K. Cho, A. Ilin, and T. Raiko. Improved learning of gaussian-bernoulli restricted boltzmann machines. *In Proc. of the 21th International Conference on Artificial Neural Networks*, pages 10–17, 2011.

[6] M. Yasuda. Monte carlo integration using spatial structure of markov random field. *Journal of the Physical Society of Japan*, 84(3):034001, 2015.

[7] M. Yasuda and K. Uchizawa. A generalization of spatial monte carlo integration. *Neural Computation*, 33(4):1037–1062, 2021.

[8] K. Sekimoto and M. Yasuda. Effective learning algorithm for restricted boltzmann machines via spatial monte carlo integration. *Nonlinear Theory and its Applications, IEICE*, 14(2):228–241, 2023.

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

[10] L. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[11] L. Mclnnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[12] D. P. Kingma and L. J. Ba. Adam: A method for stochastic optimization. *In Proc. of the 3rd International Conference on Learning Representations*, pages 1–13, 2015.