Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines

KyungHyun Cho, Alexander Ilin and Tapani Raiko

Department of Information and Computer Science
Aalto University School of Science, Finland
{firstname.lastname@aalto.fi}

Abstract. We propose a few remedies to improve training of Gaussian-Bernoulli restricted Boltzmann machines (GBRBM), which is known to be difficult. Firstly, we use a different parameterization of the energy function, which allows for more intuitive interpretation of the parameters and facilitates learning. Secondly, we propose parallel tempering learning for GBRBM. Lastly, we use an adaptive learning rate which is selected automatically in order to stabilize training. Our extensive experiments show that the proposed improvements indeed remove most of the difficulties encountered when training GBRBMs using conventional methods.

Keywords: Restricted Boltzmann Machine, Gaussian-Bernoulli Restricted Boltzmann Machine, Adaptive Learning Rate, Parallel Tempering

1 Introduction

Conventional restricted Boltzmann machines (RBM) [1, 17] define the state of each neuron to be binary, which seriously limits their application area. One popular approach to address this problem is to replace the binary visible neurons with the Gaussian ones. The corresponding model is called Gaussian-Bernoulli RBM (GBRBM) [8]. Unfortunately, training GBRBM is known to be a difficult task (see, e.g. [9, 11, 12]).

In this paper, we propose a few improvements to the conventional training methods for GBRBMs to overcome the existing difficulties. The improvements include another parameterization of the energy function, parallel tempering learning, which has previously been used for ordinary RBMs [6, 5, 3], and the use of an adaptive learning rate, similarly to [2].

2 Gaussian-Bernoulli RBM

The energy of GBRBM [8] with real-valued visible neurons $\mathbf{v}$ and binary hidden neurons $\mathbf{h}$ is traditionally defined as

$$E(\mathbf{v}, \mathbf{h}|\theta) = \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{n_h} c_j h_j, \tag{1}$$

where b_i and c_j are biases corresponding to hidden and visible neurons, respectively, W_{ij} are weights connecting visible and hidden neurons, and σ_i is the standard deviation associated with a Gaussian visible neuron v_i (see e.g. [11]).

The traditional gradient-based update rules are obtained by taking the partial derivative of the log-likelihood function $\log \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}|\theta))$, in which the hidden neurons are marginalized out, with respect to each model parameter. However, training GBRBMs even using well-defined gradients is generally difficult and takes long time (see, e.g., [11, 12]). One of the main difficulties is learning the variance parameters σ_i, which are, unlike other parameters, are constrained to be positive. Therefore, in many existing works, those parameters are often fixed to unity [9, 11, 15].

3 Improved Learning of Gaussian-Bernoulli RBM

3.1 New Parameterization of the Energy Function

The traditional energy function in (1) yields somewhat unintuitive conditional distribution in which the noise level defined by σ_i affects the conditional mean of the visible neuron. In order to change this, we use a different energy function:

$$E(\mathbf{v}, \mathbf{h}|\theta) = \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} W_{ij} h_j \frac{v_i}{\sigma_i^2} - \sum_{j=1}^{n_h} c_j h_j. \tag{2}$$

Under the modified energy function, the conditional probabilities for each visible and hidden neurons given the others are

$$p(v_i = v|\mathbf{h}) = \mathcal{N}\left(v \mid b_i + \sum_j h_j W_{ij}, \sigma_i^2\right),$$

$$p(h_j = 1|\mathbf{v}) = \text{sigmoid}\left(c_j + \sum_i W_{ij}\frac{v_i}{\sigma_i^2}\right),$$

where $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2. The update rules for the parameters are, then,

$$\nabla W_{ij} = \left\langle \frac{1}{\sigma_i^2} v_i h_j \right\rangle_{\text{d}} - \left\langle \frac{1}{\sigma_i^2} v_i h_j \right\rangle_{\text{m}}, \tag{3}$$

$$\nabla b_i = \left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{\text{d}} - \left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{\text{m}}, \tag{4}$$

$$\nabla c_j = \langle h_j \rangle_{\text{d}} - \langle h_j \rangle_{\text{m}}, \tag{5}$$

where a shorthand notations $\langle \cdot \rangle_d$ and $\langle \cdot \rangle_m$ denote the expectation computed over the data and model distributions accordingly [1].

Additionally, we use a different parameterization of the variance parameters: $\sigma_i^2 = e^{z_i}$. Since we learn log-variances $z_i = \log \sigma_i^2$, σ_i is naturally constrained to stay positive. Thus, the learning rate can be chosen with less difficulty. Under the modified

energy function, the gradient with respect to z_i is

$$\nabla z_i = e^{-z_i} \left(\left\langle \frac{1}{2}(v_i - b_i)^2 - \sum_j v_i h_j w_{ij} \right\rangle_{\mathrm{d}} - \left\langle \frac{1}{2}(v_i - b_i)^2 - \sum_j v_i h_j w_{ij} \right\rangle_{\mathrm{m}} \right).$$

3.2 Parallel Tempering

Parallel tempering (PT) learning. However, applying the same methodology to GBRBM is not straightforward: For example, a naive approach of multiplying σ_i with the temperature results in the base model with zero variances for the visible neurons, or scaling the energy function by temperature would yield infinite variances. Here, we follow the methodology of [3].

In order to overcome this problem, we propose a new scheme for constructing the intermediate models with inverse temperatures β such that

$$W_{ij}^{(t)} = \beta W_{ij}, \qquad b_i^{(t)} = \beta b_i + (1-\beta) m_i,$$
$$c_j^{(t)} = \beta c_j, \qquad \sigma_i^{(\beta)} = \sqrt{\beta \sigma_i^2 + (1-\beta) s_i^2},$$

where W_{ij}, b_i and c_j are the parameters of the current model, and m_i and s_i^2 are the overall mean and variance of the i-th visible component in the training data.

The intermediate model is thus an interpolation between the base model and the current model, where the base model consists of independent Gaussian variables fitted to the training data.

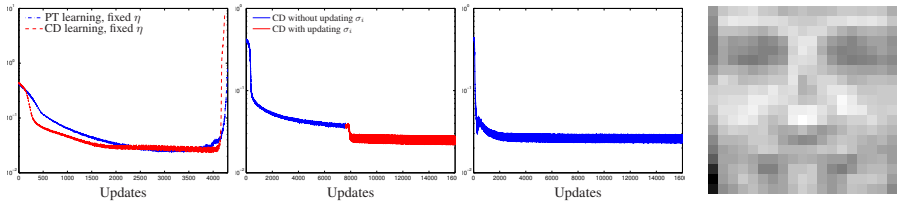
3.3 Adaptive Learning Rate

Many recent papers [2, 16, 7] point out that training RBM is sensitive to the choice of learning rate η and its scheduling. According to our experience, GBRBM tends to be even more sensitive to this choice compared to RBM. It will be shown later that, if the learning rate is not annealed towards zero, GBRBM can easily diverge in the late stage of learning.

The adaptive learning rate proposed in [2] addresses the problem of automatic choice of the learning rate. The adaptation scheme proposed there is based on an approximation of the likelihood that is valid only for small enough learning rates. In this work, we use the same adaptive learning rate strategy but we introduce an upper-bound for the learning rate so that the approximation does not become too crude.

4 Experiments

In all the experiments, we used the following settings. The weights were initialized to uniform random values between $\pm\frac{1}{n_v+n_h}$. Biases b_i and c_j were initialized to zero and variances σ_i to ones. Adaptive learning rate candidates (see [2]) were $\{0.9\eta, \eta, 1.1\eta\}$, where η is the previous learning rate. In PT learning, we used 21 equally spaced $\beta \in \{0, 0.05, \ldots, 1\}$, and in CD learning, we used a single Gibbs step.



(a) (b) (c) (d)

Fig. 1. (a)-(c): The reconstruction errors obtained by training GBRBM using a learning rate fixed to 0.001 (a), with the adaptive learning rate while updating variances from the 650-th epoch using CD learning (b) and using PT learning (c). (d): Visualization of the learned variances.

4.1 Learning Faces

The CBCL data [13] used in the experiment contains 2,429 faces and 4,548 non-faces as training set and 472 faces and 23,573 non-faces as test set. Only the faces from the training set of the CBCL data were used.

In the first experiment, we trained two GBRBMs with 256 hidden neurons using both CD and PT learning with the learning rate fixed to 0.001 while updating all parameters including σ_i^2. As can be observed from Fig. 1(a), learning diverged in both cases (CD and PT learning), which is manifested in the increasing reconstruction error. This result confirms that GBRBMs are sensitive to the learning rate scheduling. The divergence became significant when the variances decreased significantly (not shown in Fig. 1(a)), indirectly indicating that the sensitivity is related to learning the variances.

Learning Variances is Important We again trained GBRBMs with 256 hidden neurons by CD and PT learning. The upper-bound and the initial learning rate were set to 0.01 and 0.0001, respectively.

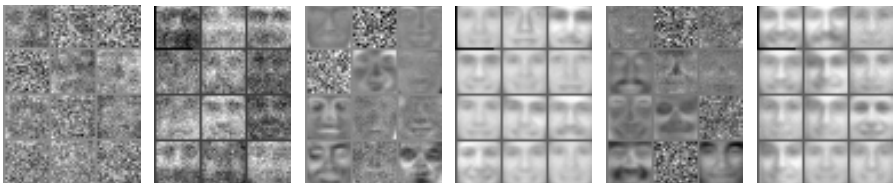
Initially, the variances of the visible neurons were not updated, but fixed to 1. The training was performed for 650 epochs. Afterwards, the training was continued for 1000 epochs, however, with updating variances.

Fig. 2(a) shows the learned filters and the samples generated from the GBRBM after the first round of training. The reconstruction error nearly converged (see the blue curve of Fig. 1(b)), but it is clear to see that both the filters and the samples are very noisy. However, the continued training significantly reduced the noise from the filters and the samples, as shown in Fig. 2(b).

From Fig. 1(b), it is clear that learning variances decreased the reconstruction error significantly. The explanation could be that the GBRBM has learned the importance, or noisiness, of pixels so that it focuses on the important ones.

The visualization of the learned variances in Fig. 1(d) reveals that important parts for modeling the face , for example, eyes and mouth, have lower variances while those of other parts are higher. Clearly, since the important parts are rather well modeled, the noise levels of corresponding visible neurons are lower.

Parallel Tempering In order to see if the proposed scheme of PT learning works well with GBRBM, an additional experiment using PT learning was conducted under the same setting, however, now updating the variances from the beginning.



(a) (b) (c)

Fig. 2. Example filters (left) and samples (right) generated by GBRBM trained using CD learning without updating variances (a), continued with updating variances (b), and trained using PT learning with updating variances from the beginning (c). 12 randomly chosen filters are shown, and between each consecutive samples 1000 Gibbs sampling steps were performed.

The observation of Fig. 1(c) suggests that learning variances from the beginning helps. It is notable that the learning did not diverge as the adaptive learning rate could anneal the learning rate appropriately.

The samples were generated from the trained GBRBM. Comparing the samples in the right figures of Fig. 2(a)–(c) suggests that the GBRBM trained using PT learning provides more variety of distinct samples, which indirectly suggests that the better generative model was learned by PT learning.

4.2 Learning Natural Images

CIFAR-10 data set [11] consists of three-channel (R, G, B) color images of size 32×32 with ten different labels.

Learning Image Patches In this experiment, the procedure proposed in [14] is roughly followed which was successfully used for classification tasks [11, 12, 4]. The procedure, first, trains a GBRBM on small image patches.

Two GBRBMs, each with 300 hidden neurons, following the modified energy function were trained on 8×8 images patches using CD and PT learning for 300 and 200 epochs, respectively.

Fig. 3 visualizes the filters learned by the GBRBMs. Apparently, the filters with the large norms mostly learn the global structure of the patches, whereas those with smaller norms tend to model more fine details. It is notable that this behavior is more obvious in the case of PT learning, whereas in the case of CD learning, the filters with the small norms mostly learned not-so-useful global structures.

The learned variances σ_i^2 of different pixels i were distributed in $[0.0308\ 0.0373]$ and $[0.0283\ 0.0430]$ in case of CD and PT learning. In both cases, they were smaller than those of the training samples s_i^2, lying between 0.0547 and 0.0697. This was expected and is desirable [11].

Classifying Natural Images The image patches were preprocessed with independent component analysis (ICA) [10] and were transformed to vectors of 64 independent components each. Then, they were used as training data for GBRBMs. GBRBMs had 200 or 300 binary hidden neurons, and were trained by persistent CD learning [18] with

(a) CD learning (b) PT learning

Fig. 3. (a) two figures visualize 128 filters with the largest norms and 128 filters with the smallest norms of the GBRBM trained using CD learning, and (b) same figures obtained from PT learning.

a fixed learning rate $\eta = 0.005$ and variances fixed to one. The minibatch of size 20 was used, and we denote this model ICA+GBRBM.

Afterwards, 49 patches were extracted from each image in a convolutional way, and the hidden activations were obtained for each patch. Those activations were concatenated to form a feature vector which was used for training a logistic regression classifier.

The best classification accuracy of 63.75% was achieved with ICA+GBRBM having 64 independent components and 300 hidden neurons after training the GBRBM for only about 35 epochs. The obtained accuracy is comparable to the accuracies from the previous research. Some of them using the variants of RBM include 63.78% by GBRBM with whitening [11], and 68.2% obtained by the mean and covariance RBM with principal component analysis [14].

Also, slightly worse accuracies were achieved when the raw pixels of the image patches were used. Using the filters obtained in the previous experiment, 55.20% (CD) and 57.42% (PT) were obtained. This suggests that it is important to preprocess samples appropriately.

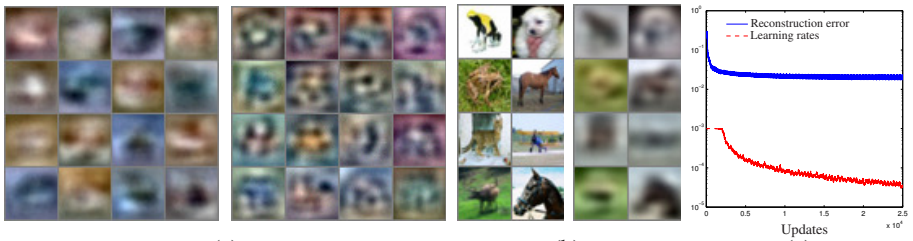
Learning Whole Images Due to the difficulty in training GBRBM, only data sets with comparably small dimensions have been mainly used in various recent papers. In case of CIFAR-10 the GBRBM was unable to learn any meaningful filters from whole images in [11].

In this experiment, a GRBM with 4000 hidden neurons was trained on whole images of CIFAR-10. It was expected that learning the variances, which became easier due to the proposed improvements, would encourage GBRBM to learn interesting interior features. CD learning with the adaptive learning rate was used. The initial learning rate and the upper-bound were set to 0.001. The training lasted for 70 epochs, and the minibatch of size 128 was used.

As shown in Fig. 4(a) the filters with the large norms tend to model the global features such as the position of the object, whereas the filters with the smaller norms model fine details, which coincides with the filters of the image patches. It is notable that the visualized filters do not possess those global, noisy filters (see Fig. 2.1 of [11]).

This visualization shows that the proposed improvements in training GBRBMs prevents the problem raised in [12] that a GBRBM easily fails to model the whole images by focusing mostly on the boundary pixels only.

Also, according to the evolution of the reconstruction error in Fig. 4(c), the learning proceeded stably. The red curve in the same plot suggests that the adaptive learning rate was able to anneal the learning rate automatically.



(a) (b) (c)

Fig. 4. (a): Two figures visualize 16 filters each with the largest norms (left) and the least norms (right) of the GBRBM trained on the whole images of CIFAR-10. (b): Two figures visualize original images (left) and their reconstructions (right). (c): The evolution of the reconstruction error and the learning rate.

Looking at Fig. 4(b), it is clear that the GBRBM was able to capture the essence of the training samples. The reconstructed images look like the blurred versions of the original ones while maintaining the overall structures. Apparently, both the boundary and the interior structure are rather well maintained.

5 Discussion

Based on the widely used GBRBM, we proposed a modified GBRBM which uses a different parameterization of the energy function. The modification led to the perhaps more elegant forms for visible and hidden conditional distributions given each other and gradient update rules.

We, then, applied two recent advances in training an RBM, PT learning and the adaptive learning rate, to a GBRBM. The new scheme of defining the tempered distributions for applying PT learning to GBRBM was proposed. The difficulty of preventing the divergence of learning was shown to be addressed by the adaptive learning rate with some practical considerations, for example, setting the upper bound of the learning rate.

Finally, the use of GBRBM and the proposed improvements were tested through the series of experiments on realistic data sets. Those experiments showed that a GBRBM and the proposed improvements were able to address the practical difficulties such as the sensitivity to the learning parameters and the inability of learning meaningful features from high dimensional data.

Despite these successful applications of GBRBM presented in this paper, training GBRBM is still more challenging than training a RBM. Further research in improving and easing the training is required.

Acknowledgements This work was supported by the summer internship and the honours programme of the department, by the Academy of Finland and by the IST Program of the European Community, under the PASCAL2 Network of Excellence. This publication only reflects the authors' views.

References

1. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines. Cognitive Science 9, 147–169 (1985)
2. Cho, K.: Improved Learning Algorithms for Restricted Boltzmann Machines. Master's thesis, Aalto University School of Science (2011)
3. Cho, K., Raiko, T., Ilin, A.: Parallel tempering is efficient for learning restricted boltzmann machines. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010). Barcelona, Spain (July 2010)
4. Coates, A., Lee, H., Ng, A.Y.: An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In: NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning (2010)
5. Desjardins, G., Courville, A., Bengio, Y.: Adaptive Parallel Tempering for Stochastic Maximum Likelihood Learning of RBMs. In: NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning (2010)
6. Desjardins, G., Courville, A., Bengio, Y., Vincent, P., Delalleau, O.: Parallel Tempering for Training of Restricted Boltzmann Machines. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 145–152 (2010)
7. Fischer, A., Igel, C.: Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines. In: Proceedings of the 20th international conference on Artificial neural networks: Part III. pp. 208–217. ICANN'10, Springer-Verlag, Berlin, Heidelberg (2010)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data with Neural Networks. Science 313(5786), 504–507 (July 2006)
9. Hinton, G.: A Practical Guide to Training Restricted Boltzmann Machines. Tech. rep., Department of Computer Science, University of Toronto (2010)
10. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley-Interscience, 1 edn. (May 2001)
11. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., Computer Science Department, University of Toronto (2009)
12. Krizhevsky, A.: Convolutional Deep Belief Networks on CIFAR-10. Tech. rep., Computer Science Department, University of Toronto (2010)
13. MIT Center For Biological and Computation Learning: CBCL Face Database #1, `http://www.ai.mit.edu/projects/cbcl`
14. Ranzato, M.A., Hinton, G.E.: Modeling pixel means and covariances using factorized third-order Boltzmann machines. In: CVPR. pp. 2551–2558 (2010)
15. Salakhutdinov, R.: LEARNING DEEP GENERATIVE MODELS. Ph.D. thesis, University of Toronto (2009)
16. Schulz, H., Müller, A., Behnke, S.: Investigating Convergence of Restricted Boltzmann Machine Learning. In: NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning (2010)
17. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. In: Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations, pp. 194–281. MIT Press, Cambridge, MA, USA (1986)
18. Tieleman, T.: Training restricted Boltzmann machines using approximations to the likelihood gradient. In: Proceedings of the 25th international conference on Machine learning. pp. 1064–1071. ICML '08, ACM, New York, NY, USA (2008)