

An Adaptive Learning Method of Restricted Boltzmann Machine by Neuron Generation and Annihilation Algorithm

Shin Kamada

Graduate School of Information Sciences,
Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami-ku,
Hiroshima, 731-3194, Japan
Email: da65002@e.hiroshima-cu.ac.jp

Takumi Ichimura

Faculty of Management and Information Systems,
Prefectural University of Hiroshima
1-1-71, Ujina-Higashi, Minami-ku,
Hiroshima, 734-8559, Japan
Email: ichimura@pu-hiroshima.ac.jp

Abstract—Restricted Boltzmann Machine (RBM) is a generative stochastic energy-based model of artificial neural network for unsupervised learning. Recently, RBM is well known to be a pre-training method of Deep Learning. In addition to visible and hidden neurons, the structure of RBM has a number of parameters such as the weights between neurons and the coefficients for them. Therefore, we may meet some difficulties to determine an optimal network structure to analyze big data. In order to evade the problem, we investigated the variance of parameters to find an optimal structure during learning. For the reason, we should check the variance of parameters to cause the fluctuation for energy function in RBM model. In this paper, we propose the adaptive learning method of RBM that can discover an optimal number of hidden neurons according to the training situation by applying the neuron generation and annihilation algorithm. In this method, a new hidden neuron is generated if the energy function is not still converged and the variance of the parameters is large. Moreover, the inactivated hidden neuron will be annihilated if the neuron does not affect the learning situation. The experimental results for some benchmark data sets were discussed in this paper.

I. INTRODUCTION

The current information technology can collect various kinds of data sets, because the recent tremendous technical advances in processing power, storage capacity, and network connected to cloud computing. Such data sample includes not only numerical values but also text such as comments, numerical evaluation such as ranking, and binary data such as pictures. Such data set is called big data. The technical methods to discover knowledge from big data are known to be a field of data mining and also developed in the research field of Deep Learning [1].

Deep Learning attracts a lot of attention in methodology research of artificial intelligence such as machine learning [2]. Especially, the industrial world is deeply impressed by the outcome to increase the capability of image processing.

©2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

The learning architecture has an advantage of not only multi-layered network structure but also pre-training. The latter characteristic means that the architecture of Deep Learning accumulates prior knowledge of the features for input patterns. Restricted Boltzmann Machine (RBM) [3] is one of popular method of Deep Learning for unsupervised learning. RBM has the capability of representing an probability distribution of input data set, and it can represent an energy-based statistical model. Moreover, the Contrastive Divergence (CD) learning procedure which is a faster algorithm of Gibbs sampling based on Markov chain Monte Carlo methods can be often used as one of the learning methods of RBM [5], [6].

The problem related to RBM is how to determine the definition of an optimal initial network structure such as the number of hidden neurons according to the features of input pattern because the traditional RBM model cannot change its network structure during learning phase. In this paper, we propose the adaptive learning method of RBM that can discover an optimal number of hidden neurons according to the training situation by applying the neuron generation and annihilation algorithm. In multi-layered neural networks, the adaptive learning method by neuron generation and annihilation algorithm during learning phase was proposed [7], [8]. The method monitors the variance of the weight vectors called the Walking Distance (WD) in the learning phase. A new neuron will be generated and inserted into the related position if the weight vector tends to fluctuate greatly even after a certain period of the training process. Moreover, the inactivated hidden neuron will be annihilated if the neuron does not affect the learning situation.

However, RBM with CD method works in an output to use binary neurons. We consider to convergence under the Lipschitz continuous condition [9]. According to [9], the energy function of RBM can be transformed to the equations under the continuous conditions with 3 kinds of parameters for visible and hidden neurons. We investigated the variance of 3 kinds parameters where energy function of RBM converges [10]. Then we selected 2 parameters which influence the

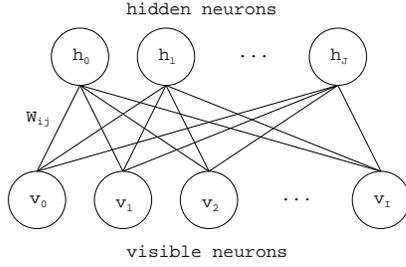


Fig. 1. The structure of RBM

convergence situation of RBM except the parameter related to input features. In this paper, we show that our proposed model has the good classification capability for the small data set (about 1000 records [11]). Moreover, we applied our proposed adaptive learning method of RBM to big data set such as CIFAR-10 [12]. From experimental results, our proposed model will be the good performance in comparison to previous RBM model [13].

The remainder of this paper is organized as follows. Section II describes the basic concept of RBM and the condition of convergence under the Lipschitz continuous is derived. In Section III-A, neuron generation and annihilation algorithm in multi-layered neural networks is explained and we apply this method into RBM in Section III-B. Section IV describes some experimental results. We give some discussions to conclude this paper in Section V.

II. RESTRICTED BOLTZMANN MACHINE

A. Overview

This section explains the basic concept of RBM [3]. As shown in Fig.1, RBM has the network structure with 2 kinds of layers where one is a visible layer for input data and the other is a hidden layer for representing the features of given data space. Each layer consists of some binary neurons. The traditional Boltzmann Machine has the connections between neurons in the same layer [14]. However RBM has no connection in the same layer. Therefore the calculation is easier than the traditional one from the viewpoint of the no interaction between neurons. The RBM learning employs to train the weights and some parameters for visible and hidden neurons till the energy function becomes to a certain small value. The trained RBM can represent a probability for the distribution of input data.

Let $v_i (0 \leq i \leq I)$ and $h_j (0 \leq j \leq J)$ be binary variable of a visible neuron and a hidden neuron, respectively. I and J are the number of visible and hidden neurons, respectively. The energy function $E(\mathbf{v}, \mathbf{h})$ for visible vector $\mathbf{v} \in \{0, 1\}^I$ and hidden vector $\mathbf{h} \in \{0, 1\}^J$ is given by Eq.(1). $p(\mathbf{v}, \mathbf{h})$ is the joint probability distribution of \mathbf{v} and \mathbf{h} as shown in Eq.(2).

$$E(\mathbf{v}, \mathbf{h}) = - \sum_i b_i v_i - \sum_j c_j h_j - \sum_i \sum_j v_i W_{ij} h_j, \quad (1)$$

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})), \quad (2)$$

$$Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h})), \quad (3)$$

where b_i and c_j are the parameters for v_i and h_j , respectively. W_{ij} is the weight between v_i and h_j . Z is the partition function which is given by summing over all possible pairs of visible and hidden vectors.

The parameters of RBM are updated by maximum likelihood estimation for $p(\mathbf{v}) = \sum_{\mathbf{h}} p(\mathbf{v}, \mathbf{h})$ which is the probability of \mathbf{v} . However, the computational elements increase exponentially because the optimal configuration for all possible pairs is required to obtain the maximum likelihood estimation. Therefore, the Contrastive Divergence (CD) learning procedure has been proposed as RBM training. CD method can be a faster algorithm of Gibbs sampling based on Markov chain Monte Carlo methods [5]. Then CD method is known to make a good performance even in a few sampling steps [6].

B. Convergence under the Lipschitz continuous condition [9]

CD method works in discrete space. Therefore, we consider the convergence situation of RBM under the Lipschitz continuous condition. Generally, the solution will be found by using machine learning if and only if the convexity and continuous conditions for an objective function are satisfied. However RBM learning with CD sampling method meets the situation that may cause the slight error and it may not satisfy the continuous condition because of the use of binary neuron. Even if the network has a small error in the initial step, but the total energy after a certain period of iterations will be fluctuated seriously.

Carlson et al. discussed the upper bounds on the log partition function for each parameter of RBM by the convexity and Lipschitz continuous [9]. The paper derived the following equations to measure the likelihood for 3 kinds of parameters $\theta = \{\mathbf{b}, \mathbf{c}, \mathbf{W}\}$.

$$E(\mathbf{v}, \mathbf{h}; \theta) = - \sum_i v_i b_i - \sum_j h_j c_j - \sum_i \sum_j v_i W_{ij} h_j, \quad (4)$$

$$\arg \min_{\theta} F(\theta) = \frac{-1}{N} \sum_{n=1}^N \log p_{\theta}(\mathbf{v}_n) = f(\theta) - g(\theta), \quad (5)$$

$$f(\theta) = \log \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}; \theta)), \quad (6)$$

$$g(\theta) = \frac{1}{N} \sum_{n=1}^N \log \sum_{\mathbf{h}} \exp(-E(\mathbf{v}_n, \mathbf{h}; \theta)), \quad (7)$$

where $\mathbf{v}_n = \{v_1, v_2, \dots, v_N\}$ is a given input data, N is the number of samples of input data. Eq.(6) and Eq.(7) are log likelihood functions for ideal model and real model for input data, respectively. $f(\theta)$ is the log partition function and it is estimated by the sampling method such as CD method. $f(\theta)$ can be transformed to Eq.(8), Eq.(9) and Eq.(10) for each parameter under the assumption of Lipschitz continuous (see [9] for details). As a result, these equations can be derived from the Taylor's expansion under the partial derivative of $f(\theta)$ in each parameter.

$$f(\{\mathbf{b}, \mathbf{c}^k, \mathbf{W}^k\}) \leq f(\theta^k) + \langle \nabla_{\mathbf{b}} f(\theta^k), \mathbf{b} - \mathbf{b}^k \rangle + \frac{I}{2} \|\mathbf{b} - \mathbf{b}^k\|_{\infty}^2, \quad (8)$$

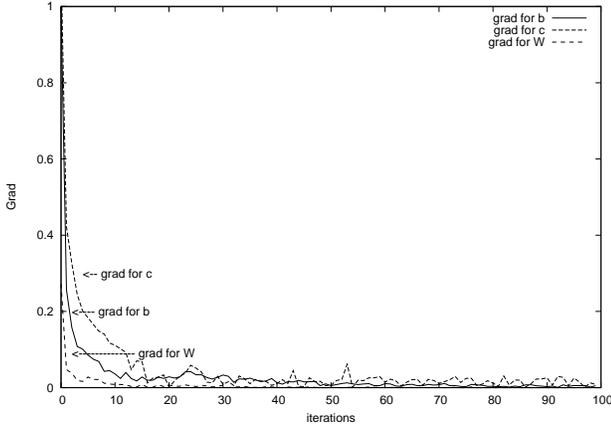


Fig. 2. Gradient for b, c and W

$$f(\{\mathbf{b}^k, \mathbf{c}, \mathbf{W}^k\}) \leq f(\theta^k) + \langle \nabla_{\mathbf{c}} f(\theta^k), \mathbf{c} - \mathbf{c}^k \rangle + \frac{J}{2} \|\mathbf{c} - \mathbf{c}^k\|_{\infty}^2, \quad (9)$$

$$f(\{\mathbf{b}^k, \mathbf{c}^k, \mathbf{W}\}) \leq f(\theta^k) + \text{tr}((\mathbf{W} - \mathbf{W}^k)^T \nabla_{\mathbf{W}} f(\theta^k)) + \frac{2IJ}{2} \|\mathbf{W} - \mathbf{W}^k\|_{S^{\infty}}^2, \quad (10)$$

where I and J are the number of visible neurons and hidden neurons, respectively. S^{∞} is Shatten-norm. $\langle \mathbf{a}, \mathbf{b} \rangle$ means an inner product between 2 vectors of \mathbf{a} and \mathbf{b} . The upper bound equations for each parameter are based on the definition of Lipschitz continuous condition, that is the third term of right side of each equation means the range of learning convergence. Therefore, the RBM learning by CD method will be converged if the variance for each parameter falls into a certain range during training.

We investigated the change of gradients for 3 kinds of parameters $\theta = \{\mathbf{b}, \mathbf{c}, \mathbf{W}\}$ for the data set MNIST [15]. Fig.2 shows the gradients for each parameter in the simulation result. As shown in Fig.2, the gradients for each parameter became gradually small. The gradient for parameter \mathbf{c} was fluctuated large in 3 kinds of parameter. Moreover, the gradient for \mathbf{W} was changed according to the relationship between \mathbf{b} and \mathbf{c} . On the other hand, the gradient for \mathbf{b} was also changed in the simulation result. However we found the gradient for parameter \mathbf{b} may be fluctuated according to the features of input patterns because parameter \mathbf{b} is the bias for input space [10]. Therefore, we selected 2 parameters which have an influence on the convergence situation of RBM except the parameter related to input data.

III. ADAPTIVE LEARNING METHOD OF RESTRICTED BOLTZMANN MACHINE

A. Walking Distance in Multi-Layered Neural Network [7], [8]

The problem related to the search of an optimal number of hidden neurons has also been considered in multi-layered neural networks. The neuron generation and annihilation algorithm during learning phase was proposed [7], [8]. Typically, a hidden neuron tries to learn input features by mapping original input data into feature vector. If a neural network does not have enough neurons to be satisfied to infer, then

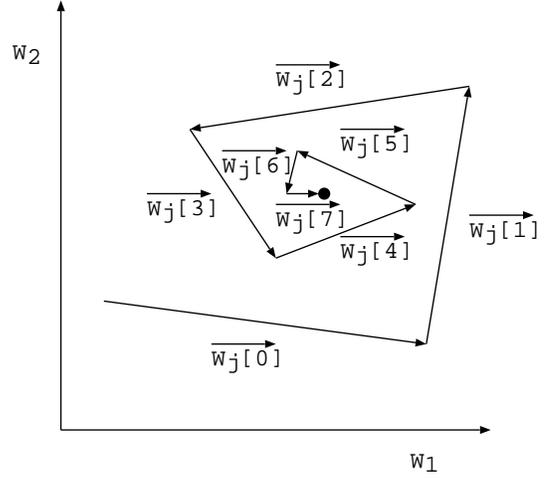


Fig. 3. An image of convergence of a weight vector

an input weight vector will tend to fluctuate greatly even after a certain period of the training process, because some hidden neurons may not represent an ambiguous patterns due to lack of the number of hidden neurons. In such a case, we can solve this problem by dividing a neuron which tries to represent the ambiguous patterns into 2 neurons by inheriting the attributes of the parent hidden neuron. The process is called the neuron generation. After an optimal number of neurons are generated, the network will be more stable within a small error. Therefore, we monitor the variance of the weight vector called the Walking Distance (WD) in hidden neurons while training as shown in Fig. 3. WD of a hidden neuron j after m training iterations is approximated by Eq.(11).

$$WD_j[m] = \gamma_w WD_j[m-1] + (1 - \gamma_w) \text{Met}(\vec{W}_j[m], \vec{W}_j[m-1]), \quad (11)$$

where $\vec{W}_j[m]$ is a weight vector of a hidden neuron j after m training iterations, Met is a metric function such as Euclidean Distance. γ_w is a constant value in $[0, 1]$. A new neuron is generated and is inserted into the neighborhood of the neuron j when WD_j is larger than the certain threshold as following equation [7], [8].

$$\Delta \varepsilon_j = \frac{\partial \varepsilon}{\partial WD_j} \cdot WD_j, \quad (12)$$

where ε is the sum of squared error of a network.

On the other hand, if a neural network has enough neurons to infer and even if an input weight vector of each neuron converges smaller than a certain value, we shall be able to find unnecessary neurons from the network. In such a case, we proposed the neuron annihilation algorithm which can annihilate the redundant neuron if the variance of output signal for a neuron is smaller than a certain threshold as following equations.

$$VA_j[m] = \gamma_v VA_j[m-1] + (1 - \gamma_v)(O_j - Act_j[m])^2, \quad (13)$$

$$Act_j[m] = \gamma_a Act_j[m-1] + (1 - \gamma_a)O_j, \quad (14)$$

where O_j is the variance of the output signal for a neuron j . γ_v and γ_a are constant values in $[0, 1]$.

B. Neuron Generation and Annihilation Algorithm in RBM

We propose the adaptive learning method of RBM that can discover an optimal number of hidden neurons by applying the neuron generation and annihilation algorithm. However, the structure of RBM has 3 kinds of parameters for visible and hidden neurons in addition to the weights between them. As mentioned in the section II-B, the RBM learning will be converged if the third term of right side in Eq.(8) - (10) become small. We consider the variance of parameters c and W except b because b is affected by many features of input patterns. Then inner product of variance of them should be monitor.

For the reason, we define the condition of neuron generation in the proposed adaptive RBM as in Eq.(15) without the gradients for b .

$$(\alpha_c \cdot dc_j) \cdot (\alpha_W \cdot dW_{ij}) > \theta_G, \quad (15)$$

where dc_j and dW_{ij} are the gradient vectors of the hidden neuron j and the weight vector i, j , respectively. α_c and α_W are the constant values for the adjustment of the range of each parameter. θ_G is an appropriate threshold value. The more θ_G is smaller, the more neuron generation is likely to be applied. A new hidden neuron is generated and is inserted into the neighborhood of the parent neuron as shown in Fig. 4(a). The initial structure of RBM should be set arbitrary neurons according to the data set before training.

If some redundant neurons are generated, these neurons may be inactivated hidden neurons that does not contribute the classification capability. The proposed adaptive learning method of RBM determines the hidden neuron that satisfies with Eq.(16), and then it annihilates the corresponding neuron as shown in Fig. 4(b).

$$\frac{1}{N} \sum_{n=1}^N p(h_j = 1 | v_n) < \theta_A, \quad (16)$$

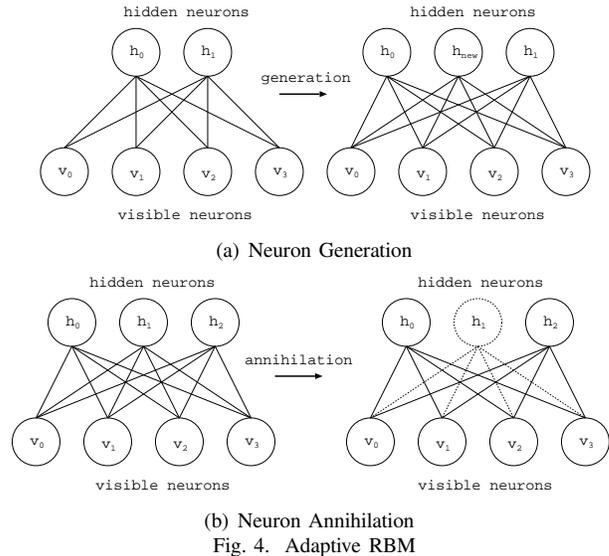
where v_n and N are same as Eq.(5). $p(h_j = 1 | v_n)$ means a conditional probability of $h_j \in \{0, 1\}$ under a given v_n . θ_A is an appropriate threshold value. The more θ_A is higher, the more neuron annihilation is likely to be applied. Fig. 4(b) shows the structure of neuron annihilation. The dot circle is the redundant neuron removed by the annihilation algorithm.

IV. EXPERIMENTAL RESULTS

This section describes experimental results to show the effectiveness of our proposed adaptive learning method of RBM.

A. Data Set

The 2 kinds of benchmark data set, ‘‘MNIST [15]’’ and ‘‘CIFAR-10 [12]’’ were used in this experiments. MNIST is popular data set of handwritten digits and has 60,000 cases for training set and 10,000 cases for test set. They are categorized into 10 classes. Each case consists of 28×28 pixels. On the other hand, CIFAR-10 is 60,000 color images data set included in 50,000 training cases and 10,000 test cases. They are categorized into 10 classes, and each case consists of 32×32 pixels. The original image data in CIFAR-10 only was preprocessed by ZCA whitening as reported in [16].



In the experiments, Pylearn2 [17], which is one of machine learning tools with libraries for Deep Learning, was used for the implementation of RBM. The following parameters were used for RBM: the training algorithms is Stochastic Gradient Descent (SGD), the batch size is 100, and the learning rate is 0.1.

B. Experimental Results

Fig. 5 shows the experimental result for MNIST. Fig. 5(a) - 5(d) show the energy curve, the gradients for each parameter, and the number of hidden neurons. In this simulation, we set parameters as the following values: the initial number of hidden neurons is 10, $\theta_G = 0.005$, $\theta_A = 0.3$. The neuron generation algorithm starts to be applied from the 10th iteration because the learning situation is fluctuated at the beginning of learning due to the selection of initial parameters. About 60 additional hidden neurons were generated until about 250 iterations as shown in Fig. 5(d). A new generated neuron inherited the weight value from the parent hidden neuron, and it was inserted into the neighborhood of the parent. The proposed RBM became smaller energy than the traditional RBM without adaptation of network structure as shown in Fig. 5(a).

After about 250 iterations, the neuron annihilation algorithm was worked as shown in Fig. 5(d), then the redundant neurons were removed from 71 to 62 gradually. The energy curve and the gradients for each parameter were not affected by the annihilation process.

Fig. 6 shows the experimental results for CIFAR-10. We set parameters as the following values for CIFAR-10: the initial number of hidden neurons is 300, $\theta_G = 0.015$, $\theta_A = 0.01$. The experimental results for CIFAR-10 showed the same characteristics as the result of MNIST. The fluctuated gradients for each parameter in the traditional RBM was higher in comparison to the result of MNIST as shown in Fig. 5(b) and Fig. 6(b) because CIFAR-10 has more ambiguous features of input patterns. On the other hand, the number of hidden neurons of our proposed RBM model was about 370 after

TABLE I
CLASSIFICATION ACCURACY ON THE MNIST

	training set	test set
traditional RBM (hidden neurons = 10)	93.4%	72.9%
adaptive RBM (hidden neurons = 77)	100.0%	83.3%

TABLE II
CLASSIFICATION ACCURACY ON THE CIFAR-10

	training set	test set
traditional RBM (hidden neurons = 300)	99.9%	70.1%
adaptive RBM (hidden neurons = 370)	99.9%	81.2%
sparse RBM (hidden neurons = 200)[13]	-	63.0%

the neuron generation. After the generation process, and the annihilation algorithm was implemented as shown in Fig. 6(d). As a result, energy curve and gradients for each parameter at last iteration were converged with smaller value than the traditional RBM as shown in Fig. 6(a) - 6(c).

In order to evaluate the classification capability with trained RBM, Hinton introduced 2 or more kinds of methods [5]. As the evaluation method for the experiment, the output layer for classification is added into the trained hidden layer of RBM, then the network between them is fine-tuned by using BP learning. Table I shows the classification accuracy for 10 kinds of images in MNIST. The classification accuracy of the proposed RBM was higher than the traditional RBM not only for training set but also for test set. Table II shows the classification accuracy for 10 kinds of images in CIFAR-10. The traditional RBM didn't classify correctly similar classes such as 'dog' and 'cat' in CIFAR-10. On the other hand, such misclassification rate was decreased in the proposed adaptive RBM because some neurons which can represent ambiguous pattern were generated. As a result, the proposed RBM showed higher classification capability compared with the result of another RBM model reported in [13] (see [13], for details for another RBM).

V. CONCLUSION

The problem related to the Deep Learning is known to be the setting for many parameters. Especially, RBM has 3 kinds of parameters for visible and hidden neurons in addition to the weights between them. In this paper, we introduced how the learning of RBM is converged under the Lipschitz continuous condition, and monitored the variance of 3 kinds of parameters during learning. Based on such observations, this paper proposed the adaptive learning method of RBM that can discover an optimal number of hidden neurons according to the training situation by applying the neuron generation and annihilation algorithm. Especially, some parameters were used for the neuron generation condition. In the simulation, we verified the effectiveness of our proposed method with 2 kinds of benchmark data set. As a result, our proposed RBM showed the higher classification capability with stable energy in comparison to the traditional RBM. In order to improve the classification capability further, the deep architecture that can represent more multiple features of input patterns with hierarchical level such as Deep Belief Network [18] is required. In future, we will develop the RBM learning method with such

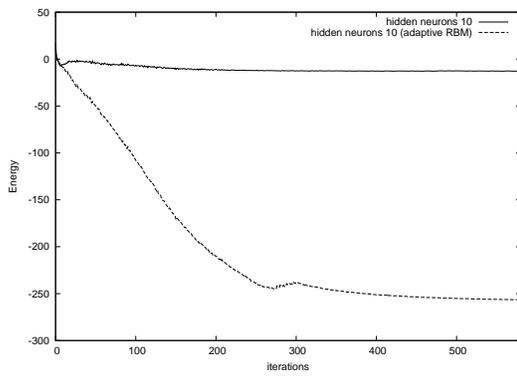
a hierarchical structure where an optimal number of hidden layers is automatically defined.

ACKNOWLEDGMENT

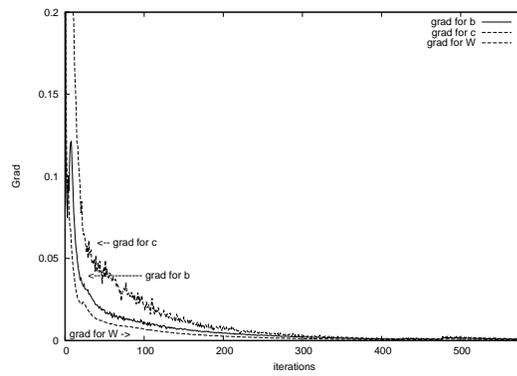
This work was supported by JSPS KAKENHI Grant Number 25330366.

REFERENCES

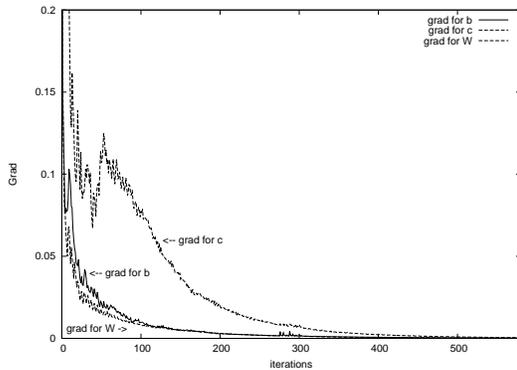
- [1] V.Le.Quoc, R.Marc's Aurelio, et.al, *Building high-level features using large scale unsupervised learning*, International Conference in Machine Learning (2012)
- [2] Y.Bengio, *Learning Deep Architectures for AI*, Foundations and Trends in Machine Learning archive, Vol.2, No.1, pp.1-127 (2009)
- [3] G.E.Hinton, *A Practical Guide to Training Restricted Boltzmann Machines*, Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science, Vol.7700, pp.599-619 (2012)
- [4] Y.Bengio, P.Lamblin, D.Popovici and H.Larochelle, *Greedy Layer-Wise Training of Deep Networks*, in Advances in Neural Information Processing Systems 19 (NIPS06), pp.153-160 (2007)
- [5] G.E.Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Computation, Vol.14, pp.1771-1800 (2002)
- [6] T.Tieleman, *Training restricted Boltzmann machines using approximations to the likelihood gradient*, Proc. of the 25th international conference on Machine learning, pp.1064-1071 (2008)
- [7] T.Ichimura, *Studies on Learning and Reasoning Methods in Neural Networks*, Ph.D. Thesis, Toin University of Yokohama (1997)
- [8] T.Ichimura and K.Yoshida Eds., *Knowledge-Based Intelligent Systems for Health Care*, Advanced Knowledge International (ISBN 0-9751004-4-0) (2004)
- [9] D.Carlson, V.Cevher and L.Carlin, *Stochastic Spectral Descent for Restricted Boltzmann Machines*, Proc. of the Eighteenth International Conference on Artificial Intelligence and Statistics, pp.111-119 (2015)
- [10] S.Kamada, T.Ichimura and Y.Fujii, *A Consideration of Convergence of Energy Function in Restricted Boltzmann Machine by Lipschitz Continuity*, Proc. of IEEE SMC Hiroshima Chapter Young Researcher Workshop 2015, pp.53-56 (2015)
- [11] S.Kamada and T.Ichimura, *A Learning Method of Adaptive Deep Belief Network by using Neuron Generation and Annihilation Algorithm*, Proc. of 17th Annual Meeting of Self-Organizing Maps, pp.12.1-6 (2016)
- [12] A.Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Master of thesis, University of Toronto (2009)
- [13] S.Dieleman and B.Schrauwen, *Accelerating sparse restricted Boltzmann machine training using non-Gaussianity measures*, Proc. of Deep Learning and Unsupervised Feature Learning (2012)
- [14] D.H.Ackley, G.E.Hinton and T.J.Sejnowski, *A Learning Algorithm for Boltzmann Machines*, Cognitive Science, 9: pp.147-169. doi: 10.1207/s15516709cog0901_7 (1985)
- [15] Y.LeCun, et.al, *THE MNIST DATABASE of handwritten digits*, <http://yann.lecun.com/exdb/mnist/>, [online] (2015)
- [16] A.Coates, A.Ng and H.Lee, *An Analysis of Single-Layer Networks in Unsupervised Feature Learning*, Journal of Machine Learning Research - Proceedings Track 15:215-223 (2011)
- [17] I.Goodfellow, David Warde-Farley, et.al., *Pylearn2: a machine learning research library*, arXiv preprint arXiv:1308.4214 (2013)
- [18] G.E.Hinton, S.Osindero and Y.Teh, *A fast learning algorithm for deep belief nets*, Neural Computation, Vol.18, No.7, pp.1527-1554 (2006)



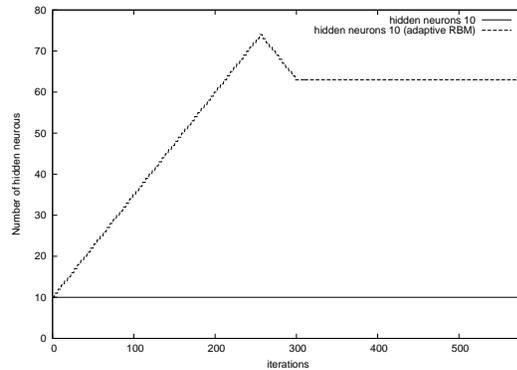
(a) Energy Function



(b) Grad for each parameter (traditional RBM)

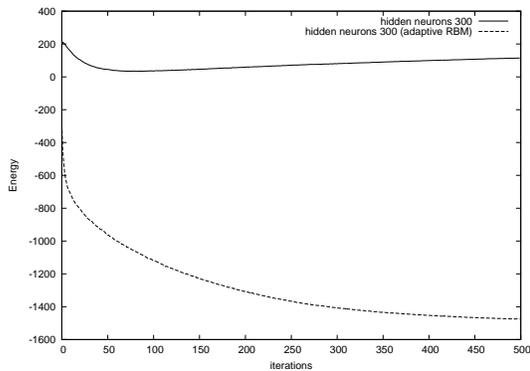


(c) Grad for each parameter (adaptive RBM)

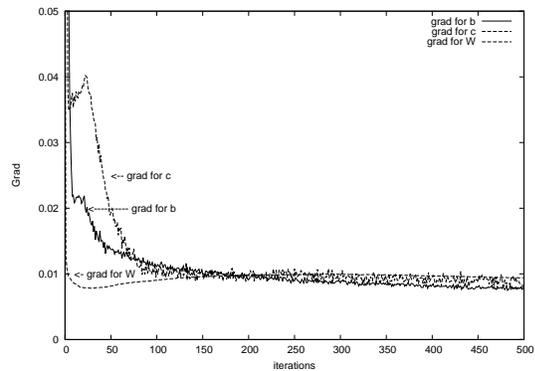


(d) No. of hidden neurons

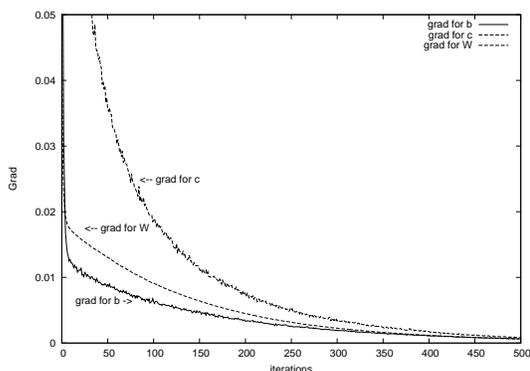
Fig. 5. MNIST (initial hidden neurons = 10)



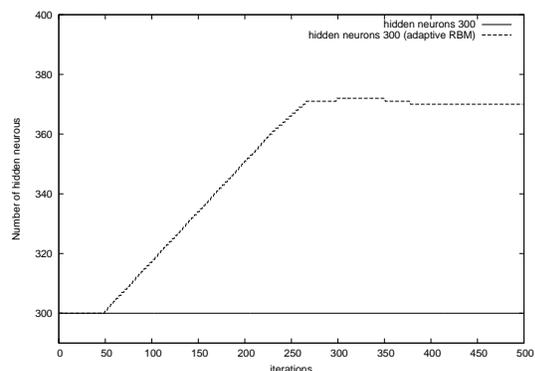
(a) Energy Function



(b) Grad for each parameter (traditional RBM)



(c) Grad for each parameter (adaptive RBM)



(d) No. of hidden neurons

Fig. 6. CIFAR-10 (initial hidden neurons = 300)