Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **1**, January **2017**

Contrastive Divergence Constrained by Reconstruction Error

Shucheng Xiao[1,a], Haijia Wu[1,b], Shan Qiu[2], Zhendong Yang[1] and Jinlan Qiao[1]

[1]Chongqing Logistical Engineering University, China;
[2] Chongqing University of Posts and Telecommunications, China.
[a]xiaosc@cqu.edu.cn, [b]wuhaijia@gmail.com

Received August, 2016; revised October, 2016

Abstract. *Deep learning is one of the most popular technology in machine learning at present. It brings a new breakthrough for intelligent information processing at Big Data era. The core algorithm of deep learning is Contrastive Divergence (CD) algorithm. The original training goal of CD is to maximize the likelihood of the probability distributions between the marginal distribution of the models visible nodes and the distribution of the training set. Aiming this training goal, the model is a one-way feature-extraction model, or encoding model. This is not suit for reconstructing information from features, or decoding model. In order to apply the model as a decoder, we need to consider the reconstruction performance additionally. A regularization constraint method based on cross entropy was designed, which adds reconstruction constraint to the visible layer of deep model. Experiments show that the new algorithm achieves reconstruction performance improvement while loses a certain likelihood degree.*
Keywords: Deep learning, Contrastive divergence, Cross entropy, Signal Reconstruction

1. **Introduction.** Deep learning [1, 2] is an effective feature extraction method in the field of Machine Learning. Deep learning apply multi-level Neural Networks to simulate the hierarchical processing mechanism of sensory perceptual system in humans brains getting an effective processing ability which is similar to the factorization therefore be able to extract feature of data by layer and learn the features in an unsupervised way. Contrastive divergence (CD) algorithm [3, 4] is the kernel of deep learning. In the original CD algorithm, the training goal is to maximize value of the marginal probability likelihood when the training samples act as the output of model, without restricting model reconstruction performance, so that the original CD algorithm trained the monodirectional model which only extracts features (encoding model) and is not applicable to the reconstruction the original data form the features (decoding model) [5]. In order to solve this problem, Bengio proposed the scheme which reconstructs the autoencoder with RBM [6, 7]. In the scheme, the training target on the likelihood of probability distribution and the reconstruction error were carried out separately, namely, target at improving the likelihood of probability distribution firstly to pre train the RBM, and then aiming at reducing the refactoring error to optimize the RBM parameters which were image connected. Bengio explanted in this article that pre training can make the model parameter adjust to the global optimum area, therefore the output will converge to the global optimal in subsequent optimization training. However, the target of pre training is not consistent with optimization training,

and it cannot guarantee the results which produced when targeting at maximum likelihood fall around the global optimal results that are solved targeting at minimizing the refactoring error [8]. In this unit, we design a visual general constraint layer framework based on cross entropy, by this framework incorporating reconstruction error constraints to the original CD algorithm, synchronize the likelihood probability distribution and the reconstruction error convergence ,so that it can improve the reconstruction performance of the model.

2. **Original CD algorithm.**

2.1. **Restricted Boltzmann Machine.** Constituting the basic unit of the deep learning web is Restricted Boltzmann Machine (RBM) [9], as shown in figure 1. It is made of m visible nodes and n hidden nodes, and visible and hidden nodes are independent of each other, which is only related to the hidden nodes, hidden nodes are also independent of each other and only related to visual node. This constraint has simplified the training of the matter. RMB for 0/1 is a random number of hidden nodes, and the visual value of node has two kinds of circumstances, including 0/1 of a random number and random number obeying Gaussian .the former is called the Bernoulli - Bernoulli RBM, latter is known as the Gaussian - Bernoulli matter.

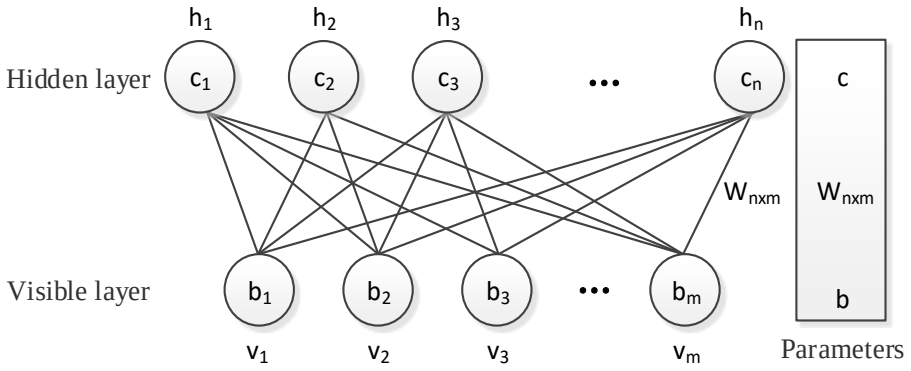


FIGURE 1. Restricted Boltzmann Machine

There are three parameters in RBM, $W_{n\times m}$ is the weighting matrix between the visual layer and hidden layer, $b = (b_1, b_2, \cdots, b_m)$ is a visible nodes offsets, $c = (c_1, c_2, \cdots, c_n)$ is hidden nodes of the offset. The three parameters determine the matter extracted from the sample of how to from a m d n characteristics.

2.2. **Extract process of the sample feature.** RBM can extracted from the sample of how to from a m d n characteristics through three parameters .Samples here in this unit, the use of RBM and the characteristics of the process [10].

For samples $x = (x_1, x_2, \cdots, x_m)$, through matter can extract the feature $y = (y_1, y_2, \cdots, y_n)$. The feature extraction of the rules is as follows:

1) Using the formula $p(h_i = 1|v) = \sigma(\Sigma_{j=1}^{m} w_{ij} \times v_j + c_i)$ to calculate the hidden layer of the probability of the ith a node value is 1, which $v_j = x_j$, $\sigma(x) = 1/(1 + e^{-x})$ known as the sigmoid function;

2) To generate a random number between 0 and 1, if the random number is less than $p(h_i = 1|v)$, the value is 1, otherwise 0.

If, in turn, the characteristics of the known samples to y, through and reconfigurable features samples. According to Bernoulli - Bernoulli RBM reconstruction process is as follows:

1) Using the formula $p(v_j = 1|h) = \sigma(\Sigma_{i=1}^{n} w_{ij} \times h_i + b_j)$ to calculate the visual layer first j a probability of node values are 1, among them $h_i = y_i$;

2) To generate a random number between 0 and 1, if the random number is less than $p(v_j = 1|h)$, the x_j is 1, otherwise 0.

For Gaussian - Bernoulli RBM refactoring process with Bernoulli - Bernoulli matter is slightly different, the specific process is as follows:

1) Making use of the formula $p(v_j|h) = N(\Sigma_{i=1}^{n} w_{ij} h_j + b_i, 1)$ to obtain the probability distribution of the first j a visible layer node, among them $h_i = y_i$, $N(a, b)$ stand for Gaussian distribution with mean variance for b;

2) Producing a rule that the Gaussian distribution of random Numbers, the averages is $\Sigma_{i=1}^{n} w_{ij} h_j + b_i$, variance is 1. x_j is the value of the random number.

RBMs training goal is to adjust the model parameters to make the edge of the probability distribution $p(v)$ of the model fitting the probability distribution $q(x)$ of training samples, the CD algorithm training aims to:

$$\arg\max_{\{W,b,c\}} \sum_{k=1}^{K} \log p(v^{(k)}) \tag{1}$$

Here the K is the size of training sample.

Through the objective function, we can make the parameters of iterative function:

$$\begin{aligned} \Delta w_{ij} &= \eta \left(\left\langle v_i^+ h_j^+ \right\rangle - \left\langle v_i^- h_j^- \right\rangle \right) \\ \Delta b_j &= \eta \left(\left\langle h_j^+ \right\rangle - \left\langle h_j^- \right\rangle \right) \\ \Delta c_i &= \eta \left(\left\langle v_i^+ \right\rangle - \left\langle v_i^- \right\rangle \right) \end{aligned} \tag{2}$$

Among them, η for the learning rate, v_i^+ stand for the training sample the i d component, h_j^+ stand for the training sample corresponding state first j d component of hidden layers, v_i^- and h_j^- respectively reconstructed visual layer state components and their corresponding state of hidden layers, $\langle \cdot \rangle$ called calculate on the average of the training sample set.

CD training process by figure 2 image description, including sample size K, I stands for visible total number of nodes, J stands for the total number of hidden nodes.

3. **Restriction method of the reconstruction of visible layer based on cross entropy.** In order to add additional constraints during the CD training process, some regularized term can be added to the objective function [11, 12]. An assignment matrix $Z \in R^{J \times K}$ will be used to control the regularized term precisely. Each element $z_i^{(k)} \in [0, 1]$ in the matrix represents the probability of assigning the kth signal to ith neuron. When corresponded to the RBM, each element indicates the probability that the kth training sample activates the ith visible node in reconstruction process. The assignment matrix is important because it provide a set of rows and columns template to the visible layers. The rows correspond to the activation performance of nodes on the visible layers when given signals of different features in the reconstruction process. And the columns correspond to the activation performance of each visible layer when given signals of special feature in the reconstruction process. The restriction that the assignment matrix put on the visible layers can be reflect through the cross entropy. Nair and Hinton [13] gave the origin definition of the cross entropy in their article. The cross entropy for the visual layers defined as:

$$z_i^{(k)} \log v_i^{(k)-} + (1 - z_i^{(k)}) \log(1 - v_i^{(k)-}) \tag{3}$$

Combined the expressions below and it turned out to be:

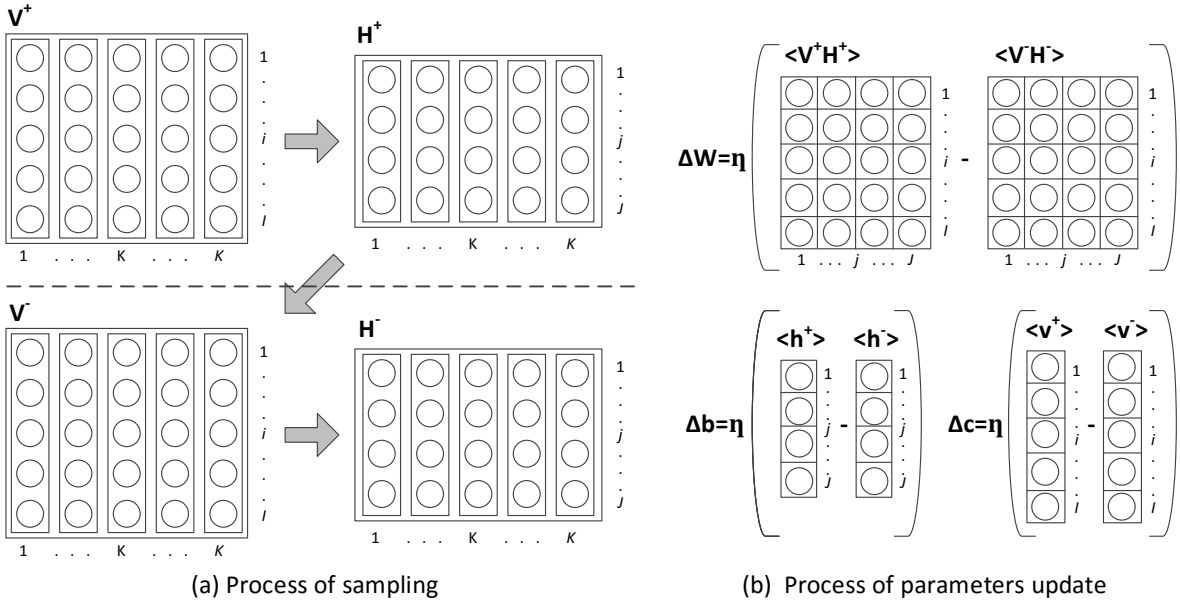


FIGURE 2. The process of CD training

$$\log\left[\left(v_i^{(k)-}\right)^{z_i^{(k)}}\left(1-v_i^{(k)-}\right)^{1-z_i^{(k)}}\right] \tag{4}$$

It can be seen from the combined form that, if reconstruction probability $v_i^{(k)-}$ of primary sampling when kth sample was given to the ith node on the visible layer resonate with assignment $z_i^{(k)}$ made by the matrix z, the cross entropy will rank its top.(Resonate means be coincident on the vector direction, or they take their maximal or minimal at the same time). So setting the cross entropy as a regularized term of the objective function in the CD training process, can control the performance of the reconstruction on the visible layers, and add constraints on the reconstruction information of visible layers [14]. The objective function with the cross entropy goes as follows:

$$\arg\max_{\{W,b,c\}}\left\{\sum_{k=1}^{K}\left[\log p(v^{(k)})+\lambda\sum_{j=1}^{J}\left(z_i^{(k)}\log v_i^{(k)-}+(1-z_i^{(k)})\log(1-v_i^{(k)-})\right)\right]\right\} \tag{5}$$

Here, λ represent the coefficient of regularized term.

The formula based on the objective function is as follows:

$$\begin{aligned}\Delta w_{ij}&=\eta\left(\left\langle v_i^+h_j^+\right\rangle-\left\langle v_i^-h_j^-\right\rangle\right)+\varepsilon\left\langle\left(v_i^--z_i\right)h_j^-\right\rangle\\&=\eta\left(\left\langle v_i^+h_j^+\right\rangle-\left\langle s_ih_j^-\right\rangle\right)\\\Delta b_j&=\eta\left(\left\langle h_j^+\right\rangle-\left\langle h_j^-\right\rangle\right)\\\Delta c_i&=\eta\left(\left\langle v_i^+\right\rangle-\left\langle v_i^-\right\rangle\right)+\varepsilon\left\langle v_i^--z_i\right\rangle\\&=\eta\left(\left\langle v_i^+\right\rangle-\langle s_i\rangle\right)\end{aligned} \tag{6}$$

Here s is called constraint matrix, $s_i^{(k)}=\phi z_i^{(k)}+(1-\phi)v_i^{(k)-}$, and ϕ is called assignment degree, $\phi=\varepsilon/\eta$.

The iteration is shown in Figure 3.

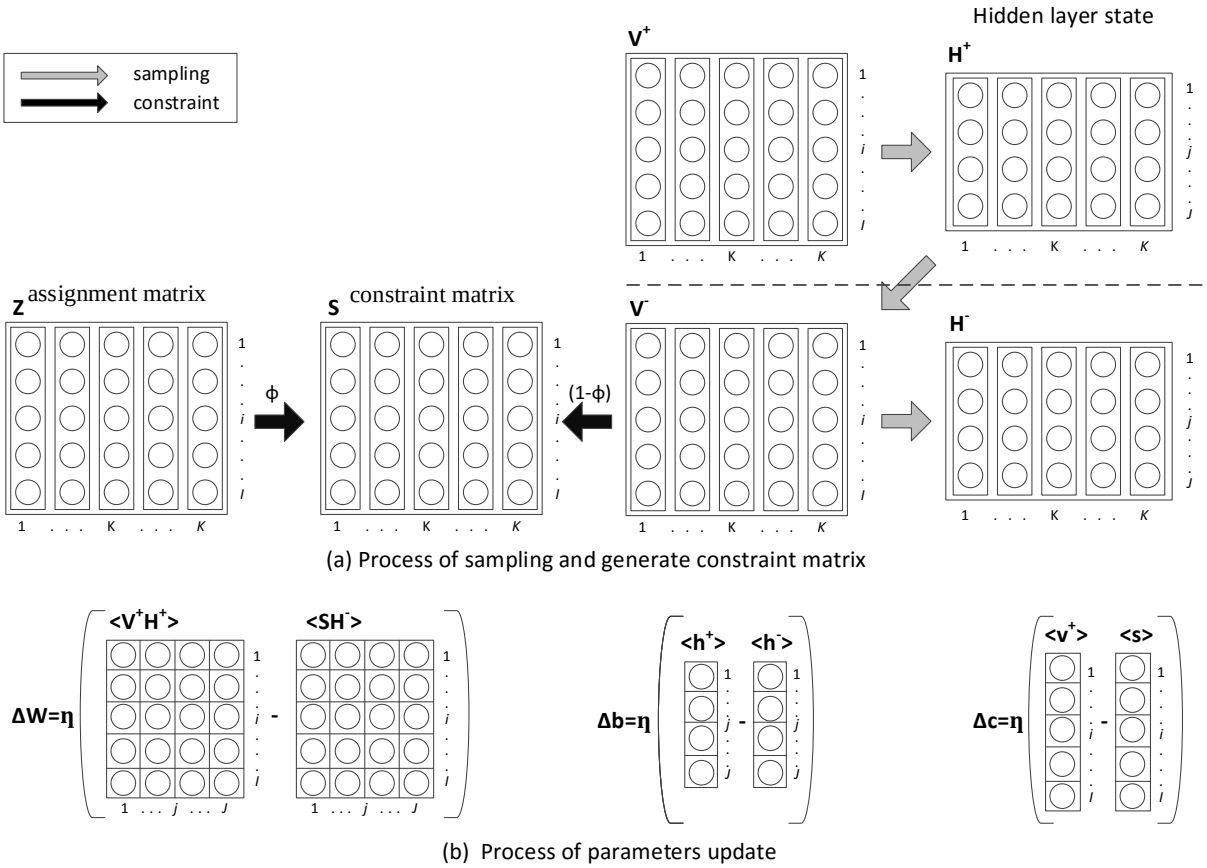


FIGURE 3. The process of CD training

Adding reconstruction constraints to the visible layers based on the frame is realized by imposing constraints on the training process through the assignment matrix. And the reconstruction constraints mean to minimize the reconstruction error $(v^+ - V^-)$ of primary sampling as small as possible. In fact, to achieve the constraints, just let the assignment matrix be $Z = V^+$. At the same time, only when the $V^- = V^+$, the cross entropy that described by expression(3) can take the maximal, and thus, reach the purpose of guiding the reconstruction value by training samples.

4. **Algorithm performance analyses.** The origin purpose of training of CD algorithm is to maximize the marginal probability distribution of model and the likelihood of the training samples distribution of. Adding additional reconstruction error may have effect on the likelihood. Analysing the algorithm should take both reconstruction performance and likelihood into consideration.

4.1. **Performance analysis design.** Considering the computational complexity, traversal statistic is applied in a small RMB to ensure the accuracy of the marginal probability distribution. In fact the computational complexity turns out to be 2^{v+h}, when counting all the status of the Bernoulli-Bernoulli RBM consisted of v visible nodes and h hidden nodes. We set v = 8, h = 6 for the test. The training sample was a 8 dimensional data set that obey the Bernoulli distribution. The data set was obtained by sampling a one dimensional Gaussian mixture density function. The methods go as follows: Firstly, generate 10 groups of data sets which match different Gaussian distributions, and the volumes of these groups are corresponding to the mixture coefficient of Gaussian mixture function. And then merge these 10 groups into a Gaussian mixture distribution data set

with 10 Gaussian components, normalize the data set and finally get the data set S. The probability distribution of the sample points in the generated training data set S is shown in Figure 4.

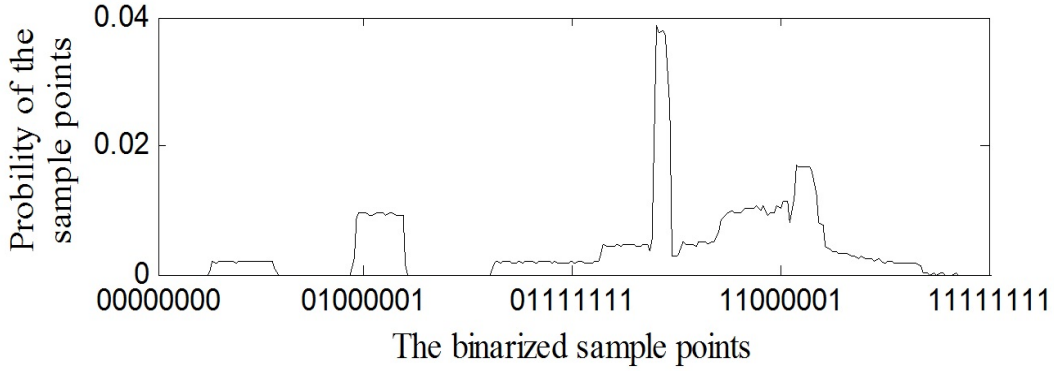


FIGURE 4. The probability distribution of the sample points in the generated training data set

4.2. **The relationship between ion of reconstruction errors and distribution of training sample.** The expectations of reconstruction error reflect the reconstructing ability of data which meet the distribution of training sample of RBM. The smaller the expectation is the stronger the better the reconstruction ability is, otherwise worse. The index's calculation method is as follows: To calculate v_i^-, encode and reconstruct all the possible input v_i^+ of visible nodes respectively, then calculate the reconstruction error $\Delta v_i = \left|v_i^+ - v_i^-\right|$. And the expectation of the errors under the distribution of the training sample named as ER.

$$E_R = E_{p(s)}(\Delta v_i) = \sum\nolimits_i p(v_i) \times \Delta v_i \tag{7}$$

The degree of assignment will influence the reconstruction error. The greater the degree is, the smaller the reconstruction error of data is at the points with more probability distribution of training sample. And the cost will be that at the points with less probability distribution of training sample, the reconstruction error will be larger. As it is shown in Figure 5(a f)

In the Figure 5 shown above, horizontal axis represents the RBM inputs while vertical axis represents the reconstruction error. Comparing with probability distribution of training sample in Figure 4, the trend of reconstruction error is completely contrary to the probability distribution of training sample. The reconstruction error is small at the points where the probability distribution is large. As the increment of assignment degree ϕ, the reconstruction error become close to zero where the probability distribution is large. And the reconstruction error is close to 8 where the probability distribution is small.(The maximal reconstruction error of Bernouli-Bernoulli RBM with 8 visible points is 8). Table 1 provide the value of reconstruction error expectation ER with different ϕ.

It can be seen from the table above, as the ϕ increase, the expectation of reconstruction error gradually reduce, $\phi = 0$ is the original CD algorithm. The trends show that after adding the reconstruction error constraints, the RBM shows better reconstruction performance to the data which meet the probability distribution of training sample.

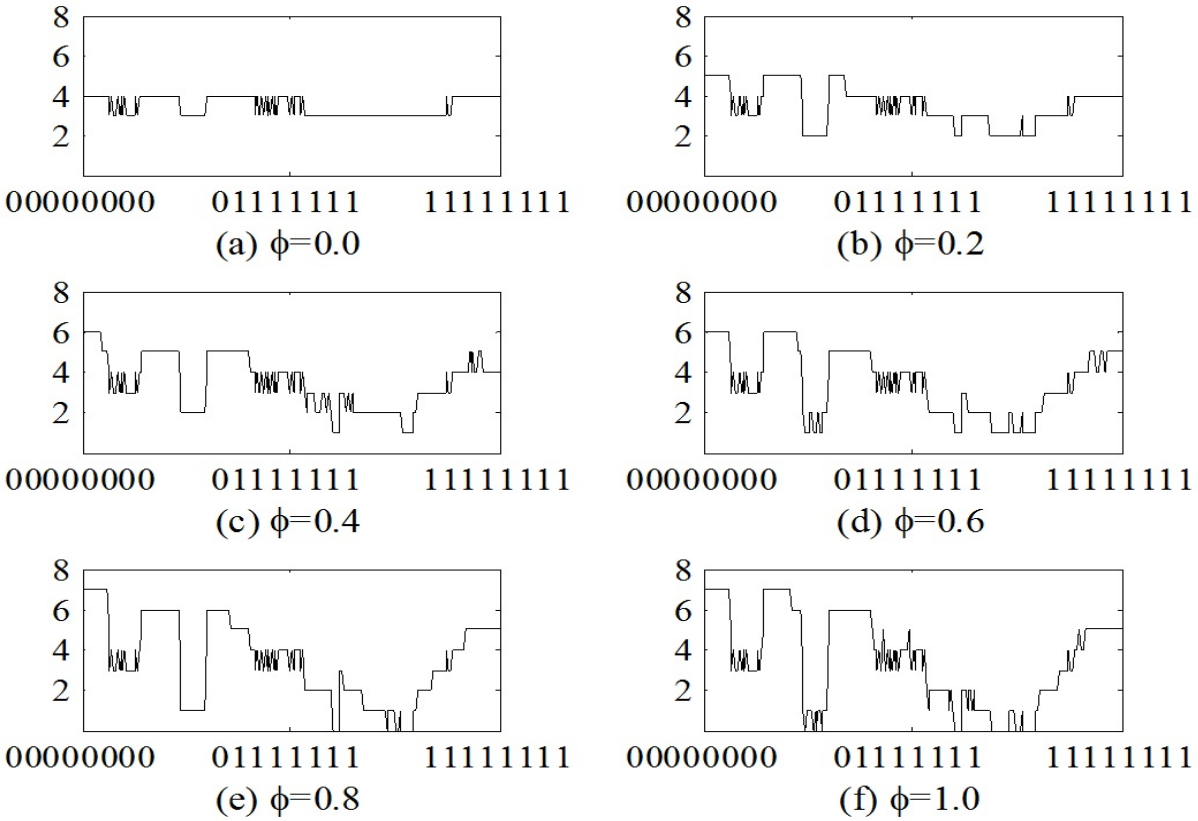


FIGURE 5. Reconstruction errors under different assignment degree

4.3. **The relationship between the expectation of reconstruction error, the distance of likelihood and the assignment degree.** The distance of likelihood shows the degree of fitting of the probability distribution of visible nodes and the training sample of the RBM. The smaller the distance is, the higher the fitting degree is, otherwise worse. To calculate the distance, traverse the v and h of RBM and construct joint distribution p(v,h), of v and h, count the marginal distribution of v, and then evaluate the likelihood by KL distance[15,16]:

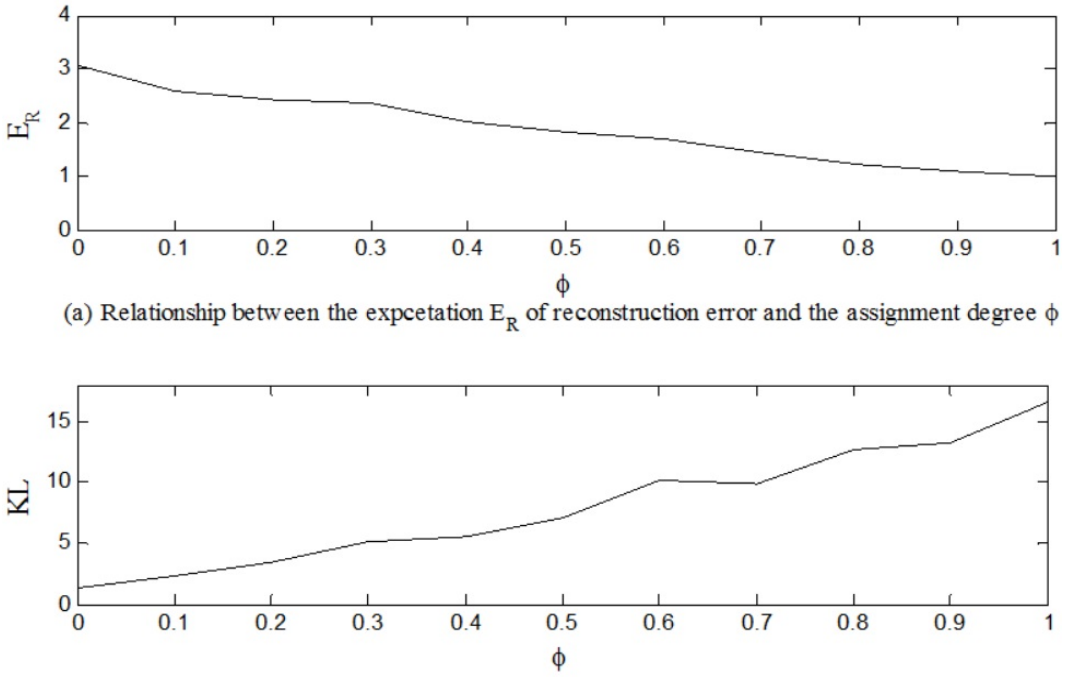
$$KL = \sum_{x \in S} p_s(x) \ln \frac{p_s(x)}{p_v(x)} \tag{8}$$

Here, S is a set of training sample. $p_s(x)$ represent the probability that x locate in the distribution of training sample. $p_v(x)$ give the probability of x on marginal distribution of the visible nodes in RBM. A small number should be add to the $p_s(x)$ and $p_v(x)$ to avoid the situation that denominator or the base of logarithm is zero. It was proved in test 1 that, in CD algorithm, adding reconstruction constraints can improve the reconstruction performance. The effect of the constraint on the probability distribution of the RBM visual node and the probability distribution of the training samples is tested below. As

TABLE 1. Expectation of Reconstruction error under different assignment degree

ϕ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
E_R	3.25	3.01	2.95	2.89	2.76	2.65	2.56	2.47	2.35	2.25	2.19

shown in Figure 6 below, the expectation R of reconstruction error and the distance of likelihood KL change along trend of the assignment degree.



(a) Relationship between the expcetation E_R of reconstruction error and the assignment degree ϕ

FIGURE 6. Comparison of the change trends of the two indexes ER and KL

It can be seen from the figure 6, with the increment of the assignment degree ϕ, the expectation ER of reconstruction error reduced gradually, the reconstruction performance improved, but as the distance of likelihood gradually increased, the likelihood reduced.

5. **Conclusion.** The goal of traditional CD algorithm is to maximize the likelihood of marginal probability distribution of the RBM visible nodes and the probability distribution of training samples. In order to apply the RBM in constructing the encoder and decoder, the performance of reconstruction should be considered. This section added the refactoring error constraints to traditional CD algorithm based on general constraint framework of cross-entropy. The results show that the reconstruction performance of CD algorithm can be improve that if additional reconstruction error constraints were put on it, but likelihood will decline relatively.

Acknowledgment. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Hinton , Y.Bengio, Learning Deep Architectures for AI, *Foundations and Trends in Machine Learning, 2009*, vol. 2, no. 1, pp. 1-127, 2009.
[2] S. Srinivas ,R. V. Babu, Deep Learning in Neural Networks, *An Overview. Computer Science*, 2015.
[3] G. E. Hinton, S. Osindero, Y. W. The, A Fast Learning Algorithm for Deep Belief Nets, *J. Neural Comp.*, vol. 28, no. 7, pp. 1527-1554, 2006.
[4] A. M. Sheri, A. Rafique, W. Pedrycz, et al., Contrastive divergence for memristor-based restricted Boltzmann machine, *J. Engineering Applications of Artificial Intelligence*, vol. 37, pp. 336-342, 2015.
[5] G. Duan,W. Hu , J.Wang, Research on the natural image super-resolution reconstruction algorithm based on compressive perception theory and deep learning model,*J. Neurocomputing*, 2016.
[6] G. Desiardins, A. Courville, Y. Bengio, Tempered Markov Chain Monte Carlo for Training of Restricted Boltzmann machine, *Proceedings of AISTATS*. pp. 145-152, 2010.

[7] C. X. Zhang , J. I. Nan-Nan , G. W. Wang, Restricted Boltzmann Machines, *Chinese Journal of Engineering Mathematics*, 2015.
[8] H. j. Wu, X. Q. zhang, M. Sun, J. B. Yang, Biasness of contrastive divergence in deep learning, *Journal of PLA University of Science and Technology (Natural Science Edition)*, vol.3, pp. 224-230, 2015.
[9] G. E.Hinton, A Practical Guide to Training Restricted Boltzmann Machines, *[J]. Momentum,* vol. 9, no. 1, pp. 599-619, 2010.
[10] N. L. Roux , Y. Bengio, Representational power of restricted boltzmann machines and deep belief networks, *[J]. Neural Computation*, vol.20, no.6, pp. 1631-1649, 2008.
[11] S. Xiao , J. Wu, E. He, et al., Identification of software NFR based on the fuzzy-QFD model, *International Journal of Security and its Applications*, vol.9, no. 11, pp. 145-154, 2015.
[12] S. Xiao, J. Wu, H. He, et al., An Emergency Logistics Transportation Path Optimization Model by Using Trapezoidal Fuzzy, The 11th international conference on the fuzzy systems and knowledge discovery (FSKD), pp. 199-203,2014.
[13] V.Nair, G.Hinton. 3D object recognition with deep belief nets, *In NIPS 2009.*
[14] P. T. D. Boer, D. Kroese, S. Mannor, et al., A tutorial on the cross-entropy method, *[J]. Annals of Operations Research*, vol. 134, no. 1, pp.19-67, 2005.
[15] B. R. Chang, H. F. Tsai, C. M. Chen, Evaluation of virtual machine performance and virtualized consolidation ratio in cloud computing system, [*J]. Journal of Information Hiding & Multimedia Signal Processing*, vol. 4, no. 3, pp.192-200, 2013.
[16] K. L. Divergence, K. Leibler, Divergence, *Alphascript Publishing*, 2010.