Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 350102, 9 pages
http://dx.doi.org/10.1155/2015/350102

Convergence Analysis of Contrastive Divergence Algorithm Based on Gradient Method with Errors

Xuesi Ma[1,2] and Xiaojie Wang[1]

[1]*Center for Intelligence Science and Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*School of Mathematic and Information Science, Henan Polytechnic University, Jiaozuo, Henan 454000, China*

Correspondence should be addressed to Xuesi Ma; maxuesi@hpu.edu.cn

Received 12 May 2015; Accepted 1 July 2015

Academic Editor: Julien Bruchon

Contrastive Divergence has become a common way to train Restricted Boltzmann Machines; however, its convergence has not been made clear yet. This paper studies the convergence of Contrastive Divergence algorithm. We relate Contrastive Divergence algorithm to gradient method with errors and derive convergence conditions of Contrastive Divergence algorithm using the convergence theorem of gradient method with errors. We give specific convergence conditions of Contrastive Divergence learning algorithm for Restricted Boltzmann Machines in which both visible units and hidden units can only take a finite number of values. Two new convergence conditions are obtained by specifying the learning rate. Finally, we give specific conditions that the step number of Gibbs sampling must be satisfied in order to guarantee the Contrastive Divergence algorithm convergence.

1. Introduction

Deep belief networks have recently been successfully applied to resolve many problems [1–5]. Restricted Boltzmann Machines (RBMs), one of important blocks of deep belief networks, have also been widely applied in many fields [2, 6–11]. The learning of RBMs and deep belief network has been an important and hot topic in machine learning research. The learning process is a parameters estimating problem. The general parameters estimating method is challenging, Hinton proposed Contrastive Divergence (CD) learning algorithm [12]. Although it has been widely used for training deep belief networks, its convergence is still not clear. Recently, more and more researchers have studied theoretical characters of CD. Bengio and Delalleau [13] proved the use of a short Gibbs chain of length k to obtain a biased estimator of the log-likelihood gradient. Akoho and Takabatake [14] gave an information geometrical interpretation of the CD learning algorithm. Sutskever and Tieleman [15] gave proofs showing CD is not the gradient of any function. It is possible to construct regularization functions that cause it to fail to converge. Yuille [16] related CD to the stochastic approximation literature and derived elementary conditions which ensure convergence (with probability 1). However, convergence conditions are relatively strict; particularly, convergence conditions are related to the model parameter which minimize the K-L divergence $D(P_0(x) \| P(x|w))$ between the empirical distribution function of the observed data $P_0(x)$ and the model $P(x|w)$.

In this paper, we study the convergence of the CD learning algorithm. By exploring the relation between the CD algorithm and the gradient method with errors, we obtain convergence conditions of CD using the convergence theorem of gradient method with errors. Our convergence conditions are more practical than those given by Yuille [16]. We also give an analysis of convergence of the CD algorithm for RBMs, especially the convergence conditions of the CD algorithm for RBMs in which both visible units and hidden units only take a finite number of values. We give two new convergence conditions by specifying the learning rate. Finally, we give the theoretical analysis of convergence conditions of the CD algorithm for RBMs and the relationship which the learning rate and the step number of Gibbs sampling must satisfy in order to guarantee the CD algorithm convergence.

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the CD algorithm. In Section 3, we firstly propose the gradient method with errors and convergence theorem of the gradient method with errors and then relate the CD algorithm to the gradient method with errors. Convergence conditions of the CD algorithm are derived. In Section 4, we give an analysis of convergence conditions of the CD algorithm for RBMs. We draw some conclusions in Section 5.

2. Contrastive Divergence Learning Algorithm

Given a probability distribution ,

$$p(x,h;w)=\frac{1}{Z(w)}e^{-E(x,h;w)}, \tag{1}$$

The marginal likelihood is

$$p(x;w)=\frac{1}{Z(w)}\sum_{h}e^{-E(x,h;w)}. \tag{2}$$

==>

$$\frac{\partial \log p(x;w)}{\partial w}=-\sum_{h}p(h;w\mid x)\frac{\partial E(x,h;w)}{\partial w}+\sum_{\tilde{x},\tilde{h}}p\left(\tilde{x},\tilde{h};w\right)\frac{\partial E\left(\tilde{x},\tilde{h};w\right)}{\partial w}. \tag{3}$$

The log-likelihood gradient algorithm can be expressed as

$$w_{t+1}-w_t=\gamma_t\left[-\sum_{h}p(h;w_t\mid x)\frac{\partial E(x,h;w_t)}{\partial w_t}+\sum_{\tilde{x},\tilde{h}}p\left(\tilde{x},\tilde{h};w_t\right)\frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right], \tag{4}$$

where γ_t denotes the learning rate at tth update.

The first term in the bracket of the right hand of (4) can be computed exactly; however, the second term (also called the expectation under the model distribution) is intractable because the calculation of $Z(w)$ is extremely difficult. In order to apply the log-likelihood gradient algorithm, we have to do alternating blocked-Gibbs sampling from the conditionals $p(x;w\mid h)$ and $p(h;w\mid x)$. This requires an infinite number of Gibbs transitions per update to fully characterize the expectation. Hinton [12] proposed a modification of the log-likelihood gradient algorithm known as CD.

The idea of k-step CD learning (CD-k) is simple: instead of approximating the second term in the log-likelihood gradient by a sample for RBM -distribution (which would require running a Markov chain until the stationary distribution is reached), a Gibbs chain is run for only k steps . The Gibbs is initialized with a training example $x^{(0)}$ of the training set and yields the sample $x^{(k)}$ after k steps. Each step t consists of sampling $h^{(t)}$ from $p(h;w\mid x^{(t)})$ and subsequently sampling $x^{(t+1)}$ from $p(x;w\mid h^{(t)})$. The gradient (3) with w of log-likelihood for one training example $x^{(0)}$ is approximated by

$$\mathrm{CD}_k\left(w,x^{(0)}\right)=-\sum_{h}p\left(h;w\mid x^{(0)}\right)\frac{\partial E\left(x^{(0)},h;w\right)}{\partial w}+\sum_{h}p\left(h;w\mid x^{(k)}\right)\frac{\partial E\left(x^{(k)},h;w\right)}{\partial w}. \tag{5}$$

The expectation of the CD algorithm can be ascribed by

$$\mathrm{ECD}_k\left(w,x^{(0)}\right)=-\sum_{h}p\left(h;w\mid x^{(0)}\right)\frac{\partial E\left(x^{(0)},h;w\right)}{\partial w}+\sum_{\tilde{x},\tilde{h}}p_k\left(\tilde{x},\tilde{h};w\right)\frac{\partial E\left(\tilde{x},\tilde{h};w\right)}{\partial w}, \tag{6}$$

where $p_k(\tilde{x},\tilde{h};w)$ is the empirical distribution function on the samples obtained by the data $x^{(0)}$ and running the Markov chain forward for k steps, $p_k(\tilde{x},\tilde{h};w)=p(x^{(k)}=\tilde{x},h^{(k)}=\tilde{h})$.

The asymptotic unbiased estimator of the parameters can be obtained by using the log-likelihood gradient algorithm; asymptotic property of the estimator of the parameters in CD-k learning is discussed in the next section.

3. Convergence of CD Algorithm

In this section, we will study the convergence of the CD learning algorithm. The CD algorithm has a similar form with the gradient method with errors. We try to relate the CD algorithm to the gradient method with errors and derive the convergence conditions of CD. For achieving it, we firstly propose the gradient method with errors and give the convergence theorem of the gradient method with errors.

3.1. Gradient Methods with Errors and Convergence Theorem. Given the optimization problem,

$$\min f(w),\quad w\in\mathscr{R}^n, \tag{7}$$

where $\mathscr{R}^n$ denotes the n-dimensional Euclidean space and $f(w):\mathscr{R}^n\to\mathscr{R}$ is a continuously differentiable function, such that for a positive constant L we have

$$\left\|\nabla f(w)-\nabla f(\overline{w})\right\|\le L\left\|w-\overline{w}\right\|\quad \forall w,\overline{w}\in\mathscr{R}^n, \tag{8}$$

where $\|w\| = (\sum_{i=1}^{n} w_i^2)^{1/2}$.

The gradient method with errors is of the following form:

$$w_{t+1} = w_t + \gamma_t (s_t + v_t), \tag{9}$$

where γ_t is a positive step-size sequence, s_t is a descent direction, and v_t is an error.

The error v_t could be deterministic or stochastic. In both cases, the gradient method has been studied in literature [17–20]. We consider that v_t is stochastic because of the CD algorithm in this paper. The gradient method with stochastic errors can be considered as stochastic approximation algorithm or stochastic approximation procedure [21, 22]. Younes [22] has analyzed the convergence of stochastic approximation procedure (SAP) and has given the almost sure convergence conditions of SAP using ODE (Ordinary Differential Equations) approach. He generated a persistent Markov chain and studied the recursion algorithm in which several iterations of the simulation procedures are performed before updating the current parameter, with the updating being done using the average of the obtained values. Bertsekas and Tsitsiklis [18] have studied the convergence of gradient method, in which the expectation of the stochastic error v_t is zero with probability 1. We present a gradient method with different stochastic errors; convergence of the gradient method is guaranteed by the following theorem.

Lemma 1. [23] *Let X_t, Y_t and Z_t be three nonnegative random variables such that $\sum_{t=0}^{\infty} Y_t < \infty$ with probability 1 and $E_t[X_{t+1}] \le X_t + Y_t - Z_t$ for all t; then, X_t converges with probability 1 and $\sum_{t=0} Z_t < \infty$*

.

Theorem 2. *Let w_t be a sequence generated by the method*

$$w_{t+1} = w_t + \gamma_t (s_t + v_t), \tag{10}$$

where γ_t is a positive step size, s_t is a descent direction, and v_t is a stochastic error. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields ($\mathcal{F}_t$ should be interpreted as the history of the algorithm up to time t). We assume the following:

(1) *$f(w) \ge 0$, $w \in \mathcal{R}^n$, $f(w_t)$, and s_t are $\mathcal{F}_t$-measurable.*

(2) *There exist positive scalars c_1 and c_2 such that*

$$\begin{aligned} c_1 \left\| f(w_t) \right\|^2 &\le -\nabla f(w_t)^T s_t, \\ \left\| s_t \right\| &\le c_2 \left(1 + \left\| \nabla f(w_t) \right\|\right). \end{aligned} \tag{11}$$

(3) *We have, for all t, and with probability 1,*

$$\begin{aligned} \left\| E[v_t \mid \mathcal{F}_t] \right\| &\le \gamma_t \left(q + p \left\| \nabla f(w_t) \right\|\right), \\ E\left[\left\| v_t \right\|^2 \mid \mathcal{F}_t \right] &\le A\left(1 + \left\| \nabla f(w_t) \right\|^2\right), \end{aligned} \tag{12}$$

where p, q and A are constants, $p \ge 0$, $q \ge 0$, $A \ge 0$.

(4) *We have*

$$\begin{aligned} \sum_{t=0}^{\infty} \gamma_t &= \infty, \\ \sum_{t=0}^{\infty} \gamma_t^2 &< \infty. \end{aligned} \tag{13}$$

Then, $f(w_t)$ converges.

Proof. With fixed two vectors w and z, let ξ be a scalar parameter, and let $g(\xi) = f(w + \xi z)$. The chain rule yields $(dg/d\xi)(\xi) = z^T \nabla f(w + \xi z)$. We have

$$\begin{aligned} f(w+z) - f(w) &= g(1) - g(0) = \int_0^1 \frac{dg}{d\xi}(\xi)\, d\xi \\ &= \int_0^1 z^T \nabla f(w + \xi z)\, d\xi \\ &\le z^T \nabla f(w) + \left| \int_0^1 z^T \nabla f(w + \xi z) - z^T \nabla f(w)\, d\xi \right| \\ &\le z^T \nabla f(w) + \|z\| \int_0^1 L\xi \|z\|\, d\xi \\ &= z^T \nabla f(w) + \frac{L}{2} \|z\|^2. \end{aligned} \tag{14}$$

We apply (14) with $w = w_t$ and $z = \gamma_t(s_t + v_t)$. We obtain

$$\begin{aligned} f(w_{t+1}) \le f(w_t) &+ \gamma_t \nabla f(w_t)^T (s_t + v_t) \\ &+ \frac{\gamma_t^2 L}{2} \left\| s_t + v_t \right\|^2. \end{aligned} \tag{15}$$

Using our assumptions, the relations $\|s_t\|^2 \le 2c_2^2(1 + \|\nabla f(w_t)\|^2)$ and $\|s_t + v_t\|^2 \le 2(\|s_t\|^2 + \|v_t\|^2)$, we have

$$\begin{aligned} f(w_{t+1}) \le f(w_t) &- c_1 \gamma_t \left\| \nabla f(w_t) \right\|^2 + \gamma_t \nabla f(w_t)^T v_t \\ &+ 2c_2^2 \gamma_t^2 L \left(1 + \left\| \nabla f(w_t) \right\|^2\right) + \gamma_t^2 L \left\| v_t \right\|^2. \end{aligned} \tag{16}$$

Taking conditional expectations with respect to $\mathscr{F}_t$ and using conditions (3) and the relation $\|\nabla f(w_t)\| \le 1 + \|\nabla f(w_t)\|^2$, we have

$$\begin{aligned} &E\left[f\left(w_{t+1}\right) \mid \mathscr{F}_t\right] \\ &\quad \le E\left[f\left(w_t\right) \mid \mathscr{F}_t\right] - c_1\gamma_t \left\|\nabla f\left(w_t\right)\right\|^2 + \gamma_t \left\|\nabla f\left(w_t\right)\right\| \\ &\qquad \cdot \left\|E\left[v_t \mid \mathscr{F}_t\right]\right\| + 2c_2^2\gamma_t^2 L\left(1 + \left\|\nabla f\left(w_t\right)\right\|^2\right) \\ &\qquad + \gamma_t^2 LE\left[\left\|v_t\right\|^2 \mid \mathscr{F}_t\right] \\ &\quad \le E\left[f\left(w_t\right) \mid \mathscr{F}_t\right] - c_1\gamma_t \left\|\nabla f\left(w_t\right)\right\|^2 \\ &\qquad + \gamma_t^2 q\left\|\nabla f\left(w_t\right)\right\| + \gamma_t^2 p\left\|\nabla f\left(w_t\right)\right\|^2 + 2c_2^2\gamma_t^2 L \\ &\qquad + 2c_2^2\gamma_t^2 L\left\|\nabla f\left(w_t\right)\right\|^2 + \gamma_t^2 LA\left(1 + \left\|\nabla f\left(w_t\right)\right\|^2\right) \\ &\quad \le E\left[f\left(w_t\right) \mid \mathscr{F}_t\right] \\ &\qquad - \gamma_t\left(c_1 - q\gamma_t - p\gamma_t - 2\gamma_t c_2^2 L - \gamma_t LA\right)\left\|\nabla f\left(w_t\right)\right\|^2 \\ &\qquad + \gamma_t^2\left(q + 2c_2^2 L + LA\right), \end{aligned} \tag{17}$$

which for sufficiently large t can be written as

$$\begin{aligned} E\left[f\left(w_{t+1}\right) \mid \mathscr{F}_t\right] &\le E\left[f\left(w_t\right) \mid \mathscr{F}_t\right] \\ &\quad - \gamma_t\beta_1\left\|\nabla f\left(w_t\right)\right\|^2 + \gamma_t^2\beta_2, \end{aligned} \tag{18}$$

where β_1 and β_2 are positive scalars.

Using the assumption that $f(w_t)$ is $\mathscr{F}_t$-measurable, we have

$$E\left[f\mathscr{F}_t\left(w_{t+1}\right)\right] \le f\left(w_t\right) - \gamma_t\beta_1\left\|\nabla f\left(w_t\right)\right\|^2 + \gamma_t^2\beta_2. \tag{19}$$

By using Lemma 1 and the assumption $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$, we see that $f(w_t)$ converges.

The theorem is proved. □

Strictly speaking, the conclusion of the theorem only holds with probability 1. For simplicity, an explicit statement of this qualification often will be omitted. We will use the theorem to derive convergence conditions of CD based on the similarity between the CD algorithm and the gradient method with errors.

3.2. Convergence of CD. In order to derive convergence conditions of the CD learning algorithm using the convergence theorem of the gradient method with errors, we have to explore the relation between the CD algorithm and the gradient method with errors. We can reconstruct the CD algorithm in the form of gradient optimization problem.

The theorem of the gradient method with errors involves four basic concepts. The first is an optimization function $f(w)$, which must be a continuously differentiable function, such that, for some constant L, we have $\|\nabla f(w) - \nabla f(\overline{w})\| \le L\|w - \overline{w}\|$. The second is the descent direction s_t. The third is the error vector v_t. The last concept is the step size γ_t; γ_t can be considered as the learning rate in the CD learning algorithm. The gradient method with errors will converge provided the conditions of Theorem 2 are satisfied.

We firstly design an optimal function for CD. Let

$$f(w) = -\log p(x; w). \tag{20}$$

We can derive convergence theorem for the CD learning algorithm through selecting appropriate s_t and v_t. Next, we give the convergence theorem of the CD learning algorithm using the convergence theorem of the gradient method with errors.

Theorem 3. *The CD learning algorithm will converge providing*

(1) $\|\nabla f(w) - \nabla f(\overline{w})\| \le L\|w - \overline{w}\|, w, \overline{w} \in \mathscr{R}^n$, *where L is a positive constant,*

(2) $\sum_{t=0}^{\infty}\gamma_t = \infty, \sum_{t=0}^{\infty}\gamma_t^2 < \infty$,

(3) $\sum_{\widetilde{x},\widetilde{h}}\|p_k(\widetilde{x},\widetilde{h}; w_t) - p(\widetilde{x},\widetilde{h}; w)\| \le c\gamma_t$, *where c is a positive constant and* $\|\partial E(\widetilde{x},\widetilde{h}; w_t)/\partial w_t\|$ *is bounded.*

Proof. The CD algorithm can be described as the form of gradient optimization problem. The CD algorithm is

$$\begin{aligned} w_{t+1} = w_t + \gamma_t \Bigg\{ &-\sum_h p\left(h; w_t \mid x^{(0)}\right)\frac{\partial E\left(x^{(0)}, h; w_t\right)}{\partial w_t} \\ &+ \sum_h p\left(h; w_t \mid x^{(k)}\right)\frac{\partial E\left(x^{(k)}, h; w_t\right)}{\partial w_t}\Bigg\}. \end{aligned} \tag{21}$$

In (21), let

$$\begin{aligned} &-\sum_h p\left(h; w_t \mid x^{(0)}\right)\frac{\partial E\left(x^{(0)}, h; w_t\right)}{\partial w_t} \\ &\quad + \sum_h p\left(h; w_t \mid x^{(k)}\right)\frac{\partial E\left(x^{(k)}, h; w_t\right)}{\partial w_t} = s_t + v_t. \end{aligned} \tag{22}$$

The CD algorithm in the form of gradient optimization problem can be described as follows:

$$w_{t+1} = w_t + \gamma_t\left(s_t + v_t\right). \tag{23}$$

In (23), let

$$\begin{aligned} s_t &= -\nabla f\left(w_t\right) \\ &= -\sum_h p\left(h; w_t \mid x^{(0)}\right)\frac{\partial E\left(x^{(0)}, h; w_t\right)}{\partial w_t} \\ &\quad + \sum_{\widetilde{x},\widetilde{h}} p\left(\widetilde{x},\widetilde{h}; w_t\right)\frac{\partial E\left(\widetilde{x},\widetilde{h}; w_t\right)}{\partial w_t}, \end{aligned} \tag{24}$$

$$\begin{aligned} v_t &= \sum_h p\left(h; w_t \mid x^{(k)}\right)\frac{\partial E\left(x^{(k)}, h; w_t\right)}{\partial w_t} \\ &\quad - \sum_{\widetilde{x},\widetilde{h}} p\left(\widetilde{x},\widetilde{h}; w_t\right)\frac{\partial E\left(\widetilde{x},\widetilde{h}; w_t\right)}{\partial w_t}. \end{aligned} \tag{25}$$

Then, (23) can be considered as the gradient method with errors.

Since $s_t = -\nabla f(w_t)$, then we have

$$-\nabla f\left(w_t\right)^T s_t = \nabla f\left(w_t\right)^T \nabla f\left(w_t\right) = \left\|\nabla f\left(w_t\right)\right\|^2. \tag{26}$$

Therefore, s_t satisfies for positive scalars c_1 and c_2:

$$\begin{aligned} c_1 \left\|f\left(w_t\right)\right\|^2 &\le -\nabla f\left(w_t\right)^T s_t, \\ \left\|s_t\right\| &\le c_2 \left(1 + \left\|\nabla f\left(w_t\right)\right\|\right). \end{aligned} \tag{27}$$

Using (25),

$$\begin{aligned} &\left\|E\left[v_t \mid \mathscr{F}_t\right]\right\| \\ &= \left\|\sum_{\tilde{x},\tilde{h}} \left[p_k\left(\tilde{x},\tilde{h};w_t\right) - p\left(\tilde{x},\tilde{h};w_t\right)\right] \frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right\|, \\ &\le \sum_{\tilde{x},\tilde{h}} \left\|p_k\left(\tilde{x},\tilde{h};w_t\right) - p\left(\tilde{x},\tilde{h};w_t\right)\right\| \cdot \left\|\frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right\|. \end{aligned} \tag{28}$$

And using the condition (3) of Theorem 3, we assume the upper bound of $\|\partial E(\tilde{x},\tilde{h};w_t)/\partial w_t\|$ is G; then, we have

$$\begin{aligned} \left\|E\left[v_t \mid \mathscr{F}_t\right]\right\| &\le \sum_{\tilde{x},\tilde{h}} \left\|p_k\left(\tilde{x},\tilde{h};w_t\right) - p\left(\tilde{x},\tilde{h};w_t\right)\right\| \\ &\quad \cdot \left\|\frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right\| \\ &\le G \sum_{\tilde{x},\tilde{h}} \left\|p_k\left(\tilde{x},\tilde{h};w_t\right) - p\left(\tilde{x},\tilde{h};w_t\right)\right\| \\ &\le cG\gamma_t. \end{aligned} \tag{29}$$

So, v_t satisfies for positive scalars p and q:

$$\left\|E\left[v_t \mid \mathscr{F}_t\right]\right\| \le \gamma_t \left(q + p\left\|\nabla f\left(w_t\right)\right\|\right). \tag{30}$$

Using the assumption that the upper bound of $\|\partial E(\tilde{x},\tilde{h};w_t)/\partial w_t\|$ is G, we have

$$\begin{aligned} \left\|v_t\right\| &= \left\|\sum_h p\left(h;w_t \mid x^{(k)}\right) \frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right. \\ &\quad \left. - \sum_{\tilde{x},\tilde{h}} p\left(\tilde{x},\tilde{h};w_t\right) \frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right\| \\ &\le \left\|\sum_h p\left(h;w_t \mid x^{(k)}\right) \frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right\| \\ &\quad + \left\|\sum_{\tilde{x},\tilde{h}} p\left(\tilde{x},\tilde{h};w_t\right) \frac{\partial E\left(\tilde{x},\tilde{h};w_t\right)}{\partial w_t}\right\| \\ &\le G\left\|\sum_h p\left(h;w_t \mid x^{(k)}\right)\right\| + G\left\|\sum_{\tilde{x},\tilde{h}} p\left(\tilde{x},\tilde{h};w_t\right)\right\| \\ &= 2G. \end{aligned} \tag{31}$$

Then, $E[\|v_t\|^2 \mid \mathscr{F}_t] \le 4G^2$. We let $A = 4G^2$; then, we have

$$E\left[\left\|v_t\right\|^2 \mid \mathscr{F}_t\right] \le A\left(1 + \left\|\nabla f\left(w_t\right)\right\|^2\right). \tag{32}$$

By using Theorem 2, we have $f(w_t)$ converging; then, the CD learning algorithm will converge.

The theorem is proved. □

We derive convergence conditions of the CD algorithm in the above theorem. It is easy to find that convergence conditions mainly include three aspects of contents. The first is the function $\log p(x;w)$ of the parameter w. The second is the learning rate of the CD learning algorithm. The third includes two terms. The first term is about the error between the empirical distribution function $p_k(\tilde{x},\tilde{h};w_t)$ and the distribution function $p(\tilde{x},\tilde{h};w_t)$; this term can be controlled by the number of Gibbs sampling. The second term is the value related to the energy function $E(\tilde{x},\tilde{h};w_t)$.

These convergence conditions derived here are different from conditions that were obtained by Yuille [16]; convergence conditions which were obtained by Yuille are related to the model parameter; our convergence conditions are not related to the model parameter. Because the task of learning is to estimate the model parameter, the model parameter is generally unknown; convergence conditions in this paper have more practical significance than convergence conditions that were obtained by Yuille.

3.3. The Learning Rate and Convergence Conditions. In convergence conditions of the CD algorithm, the condition which γ_t must satisfy is a necessary condition; it is

$$\begin{aligned} \sum_{t=0}^{\infty} \gamma_t &= \infty, \\ \sum_{t=0}^{\infty} \gamma_t^2 &< \infty. \end{aligned} \tag{33}$$

Basing the fact that $\sum_{t=1}^{\infty}(1/t) = \infty$ and $\sum_{t=1}^{\infty}(1/t^2) < \infty$, we assume $\gamma_t = 1/t$ and $\gamma_0 = 0$; then, we have the following new convergence conditions derived from Theorem 3.

Corollary 4. *The CD learning algorithm will converge providing*

(1) $\|\nabla f(w) - \nabla f(\overline{w})\| \le L\|w - \overline{w}\|$, $w, \overline{w} \in \mathcal{R}^n$, *where L is a positive constant,*

(2) $\gamma_0 = 0$, $\gamma_t = 1/t$,

(3) $\sum_{\widetilde{x},\widetilde{h}} \|p_k(\widetilde{x},\widetilde{h};w_t) - p(\widetilde{x},\widetilde{h};w_t)\| \le c\gamma_t$, *where c is a positive constant and* $\|\partial E(\widetilde{x},\widetilde{h};w_t)/\partial w_t\|$ *is bounded.*

3.4. Consistency of CD. It is clear that CD is equivalent to the Monte Carlo version of the log-likelihood gradient descent as the number the MCMC step k goes to infinity, because the empirical distribution $p_k(x,h;w)$ converges to the distribution $p(x,h;w)$. It is known that CD gives a good solution; even k is relatively small. Akoho and Takabatake [14] give an intuitive interpretation about the reason why CD can approximate well by means of information geometry. In the above sections, we study the convergence of the CD algorithm; now, we consider the consistency of the CD algorithm. If w^* is a limit point of w_t, then $f(w_t)$ converges to the finite value $f(w^*)$ by Theorem 2; then, w^* is a stationary point of f ($w^* = \arg\min f(w)$); furthermore, every limit point of w_t is a stationary point of f. It is known that CD is an approximation of the log-likelihood gradient, the convergence conditions of Theorem 3 assure the error of the approximation is small enough to make CD converging. If the convergence conditions of Theorem 3 are satisfied, CD will converge. We know the conclusions of Theorems 2 and 3 hold with probability 1. We can obtain the following conclusion: if the CD algorithm converges with probability 1, the convergence point is consistent with the stationary point of the optimal function $\log p(x;w)$, which is a local optimum in general.

4. Convergence of CD Algorithm for RBMs

In this section, we consider the convergence of the CD algorithm for RBMs. In the following, we consider the case where both visible units x and hidden units h only take a finite number of values.

4.1. Convergence Conditions of CD Algorithm for RBMs. The RBMs structure is a bipartite graph consisting of one layer of visible variables $X = (X_1, \ldots, X_m)$ and one layer of hidden variables $H = (H_1, \ldots, H_n)$. The model distribution is given by $p(x,h;w) = e^{-E(x,h;w)}/Z(w)$, where $Z(w) = \sum_{x,h} e^{-E(x,h;w)}$, and the energy function is given by

$$E(x,h;w) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}h_i x_j - \sum_{j=1}^{m} b_j x_j - \sum_{i=1}^{n} c_i h_i \tag{34}$$

with w_{ij}, c_i, b_j are real-valued parameters which are denoted by w.

In Section 3, we have already considered convergence of the CD algorithm and derived convergence theorem for the CD learning algorithm based on the convergence theorem of gradient method with errors. Now, we give the convergence theorem of the CD learning algorithm for RBMs.

Theorem 5. *The CD learning algorithms for RBMs will converge providing*

(1) $\sum_{t=0}^{\infty}\gamma_t = \infty, \sum_{t=0}^{\infty}\gamma_t^2 < \infty$,

(2) $\sum_{\widetilde{x},\widetilde{h}} \|p_k(\widetilde{x},\widetilde{h};w_t) - p(\widetilde{x},\widetilde{h};w_t)\| \le c\gamma_t$, *where c is a positive constant.*

Proof. To prove the convergence of the CD learning algorithm for RBMs, we have to prove the CD algorithm satisfying three conditions of Theorem 3. Obviously, we can obtain the second condition of Theorem 3 from the first condition. Next, we prove the CD algorithm satisfying other two conditions of Theorem 3.

Firstly, we prove $f(w)$ satisfying the first condition of Theorem 3.

Since $f(w) = -\log p(x;w)$, then

$$\begin{aligned}\|\nabla f(w) - \nabla f(\overline{w})\| \le{}& \left\| \sum_h p(h;w \mid x)\frac{\partial E(x,h;w)}{\partial w} - \sum_h p(h;\overline{w} \mid x)\frac{\partial E(x,h;\overline{w})}{\partial \overline{w}} \right\| \\ &+ \left\| \sum_{\widetilde{x},\widetilde{h}} p\left(\widetilde{x},\widetilde{h};w\right)\frac{\partial E\left(\widetilde{x},\widetilde{h};w\right)}{\partial w} - \sum_{\widetilde{x},\widetilde{h}} p\left(\widetilde{x},\widetilde{h};\overline{w}\right)\frac{\partial E\left(\widetilde{x},\widetilde{h};\overline{w}\right)}{\partial \overline{w}} \right\|.\end{aligned} \tag{35}$$

Using (34), $\partial E(x,h;w)/\partial w$ is an independent variable with w since $E(x,h;w)$ is affine with w. For convenience, let

$$\frac{\partial E(x,h;w)}{\partial w} = g(x,h). \tag{36}$$

Then,

$$\begin{aligned}&\left\| \sum_h p(h;w \mid x)\frac{\partial E(x,h;w)}{\partial w} - \sum_h p(h;\overline{w} \mid x)\frac{\partial E(x,h;\overline{w})}{\partial \overline{w}} \right\| \\ &\le \left\| \sum_h g(x,h)\frac{\partial p\left(h;w' \mid x\right)}{\partial w'} \right\| \cdot \|w - \overline{w}\| \\ &\le \sum_h \|g(x,h)\| \cdot \left\| \frac{\partial p\left(h;w' \mid x\right)}{\partial w'} \right\| \cdot \|w - \overline{w}\|,\end{aligned} \tag{37}$$

where $w' = \overline{w} + \theta(w - \overline{w})$, $0 < \theta < 1$, and

$$\left\| \sum_{\widetilde{x},\widetilde{h}} p\left(\widetilde{x},\widetilde{h};w\right) \frac{\partial E\left(\widetilde{x},\widetilde{h};w\right)}{\partial w} - \sum_{\widetilde{x},\widetilde{h}} p\left(\widetilde{x},\widetilde{h};\overline{w}\right) \frac{\partial E\left(\widetilde{x},\widetilde{h};\overline{w}\right)}{\partial \overline{w}} \right\| \leq \sum_{\widetilde{x},\widetilde{h}} \left\| g\left(\widetilde{x},\widetilde{h}\right) \right\| \cdot \left\| \frac{\partial p\left(\widetilde{x},\widetilde{h};w''\right)}{\partial w''} \right\| \cdot \|w - \overline{w}\|, \tag{38}$$

where $w'' = \overline{w} + \eta(w - \overline{w})$, $0 < \eta < 1$.

Since x and h only take a finite number of values, then $\|g(x,h)\|$ has the upper bound; we assume the upper bound is G.

Since $p(x,h;w) = (1/Z(w))e^{-E(x,h;w)} = e^{-E(x,h;w)} / \sum_{x,h} e^{-E(x,h;w)}$, we have

$$\begin{aligned} \left\| \frac{\partial p(x,h;w)}{\partial w} \right\| &= \left\| -\frac{e^{-E(x,h;w)}}{Z(w)} \frac{\partial E(x,h;w)}{\partial w} + \frac{e^{-E(x,h;w)}}{Z(w)} \frac{\sum_{x,h} e^{-E(x,h;w)} \left(\partial E(x,h;w)/\partial w\right)}{Z(w)} \right\| \\ &= \left\| -p(x,h;w) \frac{\partial E(x,h;w)}{\partial w} + p(x,h;w) \sum_{x,h} p(x,h;w) \frac{\partial E(x,h;w)}{\partial w} \right\| \\ &\leq \left\| p(x,h;w) \frac{\partial E(x,h;w)}{\partial w} \right\| + \left\| p(x,h;w) \sum_{x,h} p(x,h;w) \frac{\partial E(x,h;w)}{\partial w} \right\| \\ &\leq G \left\| p(x,h;w) \right\| + G \left\| \sum_{x,h} p(x,h;w) \right\| \leq 2G. \end{aligned} \tag{39}$$

Let $G'' = 2G$; then, $\|\partial p(x,h;w'')/\partial w''\|$ has the upper bound G''.

Since

$$p(h;w \mid x) = \frac{p(x,h;w)}{p(x;w)} = \frac{e^{-E(x,h;w)}}{Z(w)} \frac{Z(w)}{\sum_h e^{-E(x,h;w)}}, \tag{40}$$

for a similar reason, $\|\partial p(h;w' \mid x)/\partial w'\|$ has the upper bound G'. Using inequalities (35), (37), and (38), we have

$$\begin{aligned} &\left\| \nabla f(w) - \nabla f(\overline{w}) \right\| \\ &\leq \sum_h G \cdot G' \cdot \|w - \overline{w}\| + \sum_{\widetilde{x},\widetilde{h}} G \cdot G'' \cdot \|w - \overline{w}\| \\ &\leq \left(\sum_h G \cdot G' + \sum_{\widetilde{x},\widetilde{h}} G \cdot G'' \right) \|w - \overline{w}\|. \end{aligned} \tag{41}$$

Let $\sum_h G \cdot G' + \sum_{\widetilde{x},\widetilde{h}} G \cdot G'' = L$; then, we have

$$\left\| \nabla f(w) - \nabla f(\overline{w}) \right\| \leq L \|w - \overline{w}\|. \tag{42}$$

□

Secondly, since $\|\partial E(\widetilde{x},\widetilde{h};w_t)/\partial w_t\|$ has the upper bound, then the condition (3) of Theorem 3 is satisfied.

By using the Theorem 3, we see that the CD learning algorithm converges.

The theorem is proved.

We obtain convergence conditions of the CD learning algorithm for RBMs. Next, we study the relationship between the learning rate and convergence conditions. Basing the fact that $\sum_{t=1}^{\infty}(1/t) = \infty$ and $\sum_{t=1}^{\infty}(1/t^2) < \infty$, we also assume $\gamma_t = 1/t$ and $\gamma_0 = 0$; then, we have the following new convergence conditions derived from Theorem 5.

Corollary 6. *The CD learning algorithm will converge providing*

(1) $\gamma_0 = 0$, $\gamma_t = 1/t$,

(2) $\sum_{\widetilde{x},\widetilde{h}} \|p_k(\widetilde{x},\widetilde{h};w_t) - p(\widetilde{x},\widetilde{h};w)\| \leq c\gamma_t$, *where c is a positive constant.*

The result of Corollary 6 shows that the convergence of the CD algorithm is related to the errors between the empirical distribution function $p_k(\widetilde{x},\widetilde{h})$ and the distribution function $p(\widetilde{x},\widetilde{h})$ providing the learning rate is deterministic; the error can be controlled by the number of Gibbs sampling.

4.2. Theoretical Analysis of Convergence Conditions. In Section 4.1, we have given the convergence conditions of the CD algorithm for RBMs; the most important term is the error between the empirical distribution function $p_k(\widetilde{x},\widetilde{h};w_t)$ and the distribution function $p(\widetilde{x},\widetilde{h};w_t)$; the empirical distribution function $p_k(\widetilde{x},\widetilde{h};w_t)$ is the empirical distribution function on the samples obtained by the data x and running the Markov chain forward for k steps, the distribution function $p(\widetilde{x},\widetilde{h};w_t)$ is the limit distribution of the empirical distribution. Fischer and Igel [24] have given the bound of the bias of $p_k(\widetilde{x},\widetilde{h};w_t)$ and $p(\widetilde{x},\widetilde{h};w_t)$:

$$\sum_{\widetilde{x},\widetilde{h}} \left\| p_k\left(\widetilde{x},\widetilde{h};w_t\right) - p\left(\widetilde{x},\widetilde{h};w_t\right) \right\| \leq \left(1 - e^{-(m+n)a}\right)^k, \tag{43}$$

where k is the step number of Gibbs sampling in tth update, m, n are the number of visible and hidden variables, and

$$a = \max\left\{ \max_{l\in\{1,\ldots,m\}} \vartheta_l, \max_{l\in\{1,\ldots,n\}} \zeta_l \right\}, \tag{44}$$

where

$$\begin{aligned} \vartheta_l &= \max\left\{ \left| \sum_{i=1}^{n} I_{\{w_{il}>0\}} w_{il} + b_l \right|, \left| \sum_{i=1}^{n} I_{\{w_{il}<0\}} w_{il} + b_l \right| \right\}, \\ \zeta_l &= \max\left\{ \left| \sum_{j=1}^{m} I_{\{w_{lj}>0\}} w_{lj} + c_l \right|, \left| \sum_{j=1}^{m} I_{\{w_{lj}<0\}} w_{lj} + c_l \right| \right\}. \end{aligned} \tag{45}$$

Then, we can draw new convergence conditions of the CD algorithm using the conclusion.

Theorem 7. *The CD learning algorithm for RBMs will converge providing*

(1) $\sum_{t=0}^{\infty} \gamma_t = \infty, \sum_{t=0}^{\infty} \gamma_t^2 < \infty$,

(2) $k \geq \log(c\gamma_t)/\log(1 - e^{-(m+n)a})$, *where* m, n *are the number of visible and hidden variables and* a *is defined in equality (43),* k *is the step number of Gibbs sampling in* t*th update* ($k \in N^+$), *and* c *is a positive constant.*

Proof. In order to prove Theorem 7, we have to prove that we can derive the condition (2) of Theorem 5 from the condition (2) of Theorem 7. Now, we prove it.

Using the inequality (43), we have

$$\sum_{\tilde{x},\tilde{h}} \left\| p_k\left(\tilde{x},\tilde{h};w_t\right) - p\left(\tilde{x},\tilde{h};w_t\right)\right\| \leq \left(1 - e^{-(m+n)a}\right)^k. \quad (46)$$

By the condition (2) of Theorem 7, we have $c\gamma_t \geq (1 - e^{-(m+n)a})^k$; using the above the inequality, we have

$$c\gamma_t \geq \sum_{\tilde{x},\tilde{h}} \left\| p_k\left(\tilde{x},\tilde{h};w_t\right) - p\left(\tilde{x},\tilde{h};w_t\right)\right\|. \quad (47)$$

The proof of the theorem is completed. □

Theorem 7 gives new convergence conditions; the second condition of the theorem is the relationship which the step number of Gibbs sampling and the learning rate must satisfy in every step of parameter updating in order to guarantee the CD algorithm convergence.

5. Conclusions

In this paper, we have studied the convergence of the CD learning algorithm. Firstly, we relate the CD learning algorithm to the gradient method with errors and give convergence conditions which ensure that the CD learning algorithm converges. Convergence conditions mainly include three aspects of contents. Our convergence conditions are different from conditions that were obtained by Yuille [16]; our convergence conditions have more practical value. Moreover, we have studied convergence of the CD algorithm for RBMs; particularly, we give convergence conditions of the CD algorithm for RBMs in which both visible units and hidden units only take a finite number of values. We give the analysis of the consistency of the CD algorithm; meanwhile, we also give two new convergence conditions by specifying the learning rate.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work is supported by the National Science Foundation of China (nos. 61273365, 11407776) and National High Technology Research and Development Program of China (no. 2012AA011103).

References

[1] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[2] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[3] A.-R. Mohamed, T. N. Sainath, G. E. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, pp. 5060–5063, May 2011.

[4] V. Nair and G. E. Hinton, "3D object recognition with deep belief nets," in *Proceedings of the Neural Information Processing Systems Conference (NIPS '09)*, pp. 1339–1347, 2009.

[5] M. A. Salama, A. E. Hassanien, and A. A. Fahmy, "Deep Belief Network for clustering and classification of a continuous data," in *Proceedings of the 10th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '10)*, pp. 473–477, December 2010.

[6] F. Feng, R. Li, and X. Wang, "Deep correspondence restricted boltzmann machine for cross-modal retrieval," *Neurocomputing*, vol. 154, pp. 50–60, 2015.

[7] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '11)*, pp. 5884–5887, May 2011.

[8] V. Mnih, H. Larochelle, and G. E. Hinton, "Conditional restricted Boltzmann machines for structured output prediction," in *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI '11)*, F. G. Cozman and A. Pfeffer, Eds., p. 514, AUAI Press, 2011.

[9] A.-R. Mohamed and G. E. Hinton, "Phone recognition using restricted boltzmann machines," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 4354–4357, IEEE, Dallas, Tex, USA, March 2010.

[10] R. R. Salakhutdinov and G. E. Hinton, "Replicated soft-max: an undirected topic model," in *Advances in Neural Information Processing Systems (NIPS 2009)*, 2009.

[11] R. R. Salakhutdinov, A. Mnih, and G. E. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 791–798, ACM, Corvallis, Ore, USA, June 2007.

[12] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[13] Y. Bengio and O. Delalleau, "Justifying and generalizing contrastive divergence," *Neural Computation*, vol. 21, no. 6, pp. 1601–1621, 2009.

[14] S. Akoho and K. Takabatake, "Information geometry of contrastive divergence," in *Proceedings of the International Conference on Information Theory and Statistical Learning (ITSL '08)*, pp. 3–9, Las Vegas, Nev, USA, July 2008.

[15] I. Sutskever and T. Tieleman, "On the convergence properties of contrastive divergence," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS '10)*, pp. 473–477, 789–795, May 2010.

[16] A. Yuille, "The convergence of contrastive divergences," in *Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS '04)*, pp. 1593–1600, December 2004.

[17] D. P. Bertsekas, Ed., *Nonlinear Programming*, Athena Scientic, Belmont, Mass, USA, 1995.

[18] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[19] V. S. Borkar, "Asynchronous stochastic approximations," *SIAM Journal on Control and Optimization*, vol. 36, no. 3, pp. 840–851, 1998.

[20] G. Pflug, "Optimization of stochastic models," in *The Interface between Simulation and Optimization*, Kluwer, Boston, Mass, USA, 1996.

[21] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[22] L. Younes, "On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates," *Stochastics and Stochastics Reports*, vol. 65, no. 3-4, pp. 177–228, 1999.

[23] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Process*, Oxford University Press, 2001.

[24] A. Fischer and C. Igel, "Bounding the bias of contrastive divergence learning," *Neural Computation*, vol. 23, no. 3, pp. 664–673, 2011.